Qualité, Fouille et Gestion des Connaissances

Fabrice Guillet

Ecole Centrale de Nantes Université de Nantes Ecole des Mines de Nantes

Ecole Doctorale STIM
Sciences et Technologies de l'Information et des Matériaux
Année 2006
N° E.D. : 366–256

HABILITATION A DIRIGER DES RECHERCHES
Spécialité : **INFORMATIQUE**

Présentée par

Fabrice GUILLET

le 8 Décembre 2006
à l'Ecole polytechnique de l'université de Nantes

Qualité, Fouille et Gestion de Connaissances

Jury

Rapporteurs :	Gilbert Ritschard, Professeur à l'université de Genève,
	Christel Vrain, Professeur à l'université d'Orléans,
	Djamel Zighed, Professeur à l'université de Lyon 2,
Examinateurs :	Jean-Pierre Barthélemy, Professeur à l'ENST de Bretagne,
	Rose Dieng-Kuntz, Directrice de Recherche à l'INRIA, Sophia Antipolis,
	Howard Hamilton, Professeur à l'université de Regina, Canada,
	Einoshin Suzuki, Professeur à l'université de Kyushu, Japon,
Invités :	Pascale Kuntz, Professeur à l'Ecole polytechnique de l'université de Nantes,
	José Martinez, Professeur à l'Ecole polytechnique de l'université de Nantes

Directeur de HdR : Henri Briand
Laboratoire : Laboratoire d'Informatique de Nantes Atlantique
2, rue de la Houssinière – BP 92208
44322 Nantes CEDEX 3.

A tous les proches qui me sont chers
et m'ont soutenu.

A mon fils, Théodore.

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction

L'ensemble des travaux synthétisé dans ce document s'insère principalement dans deux domaines de recherches : l'Extraction de Connaissances dans les Données, et la Gestion et l'Ingénierie des Connaissances. Ils sont fédérés autour d'un objectif commun, celui du traitement des *connaissance*s, et partagent une préoccupation sous-jacente pour l'*aide à la décision* et pour le développement d'*applications*.

Extraction de Connaissances dans les Données et règles d'association

Problématique

L'*Extraction de Connaissances dans les Données* (ECD), ou Knowledge Discovery in Databases (KDD), souvent mentionnée sous de terme de *Fouille de Données* (Data Mining), est un domaine de recherche multidisciplinaire, né dans les années 80 avec l'émergence des bases de données volumineuses. Il a été identifié par le MIT comme l'une des 10 technologies émergentes du 21ème siècle[1], et reconnu par l'ACM à travers la création d'un groupe de recherche international (SIG–KDD) animé par G. Piatetsky-Shapiro. L'ECD a pour objectif l'extraction de connaissances auparavant inconnues et potentiellement utiles au sein de grands volumes de données : "Knowledge Discovery is the non-trivial extraction of implicite, previously unknown, and potentially useful information from data [75]". Ce domaine de recherche par essence multidisciplinaire s'est appuyé à l'origine sur les bases de données, puis a rapidement nécessité une étroite coopération avec l'apprentissage au-

[1] *MIT Technology Review*, 2001.

tomatique, les statistiques et l'analyse de données, la visualisation, et l'aide à la décision.

L'ECD propose une méthodologie fondée sur un processus de transformation des données vers les connaissances, qui se décompose en trois étapes majeures :

1. La localisation, la sélection et le prétraitement des gisements de données ;
2. La découverte des connaissances à l'aide modèles prédictifs ou descriptifs, supervisés ou non supervisés ;
3. Le post-traitement des connaissances découvertes et leur validation par un décideur/expert des données.

Nos travaux de recherche se situent sur les phases aval (2 et 3) du processus d'ECD, au plus proche du décideur. Plus particulièrement, nous nous sommes intéressés au modèle des règles d'association (RA) qui permet de découvrir, sans connaissances préalables, des tendances implicatives au sein des données. Cet avantage d'une découverte non supervisée permise par les RA, est contrebalancé par un inconvénient majeur, celui de délivrer une quantité prohibitive de règles, qui nécessite une phase de post-traitement adaptée afin de devenir intelligible à un décideur. De notre point de vue, une solution réside dans l'hybridation de trois approches envisagées séparément dans la littérature :

1. Mesures de qualité : pour sélectionner les règles potentiellement intéressantes à l'aide de mesures adaptées (interestingness measures), et par l'élimination des redondances ;
2. Représentations graphiques : pour permettre la visualisation des règles potentiellement intéressantes ;
3. Interactivité : pour permettre à l'utilisateur d'exprimer des contraintes sur les règles afin de cibler celles qui l'intéressent.

Nous avons ainsi proposé des méthodes d'élimination des règles redondantes et de réduction des variables. Puis, nous avons étudié et comparé les mesures de qualité de règles, et avons développé des mesures originales. Enfin, considérant le problème selon une perspective d'aide à la décision dans la quelle les représentations graphiques intelligibles et l'interactivité avec l'utilisateur jouent un rôle majeur, nous avons proposé deux solutions originales et *anthropocentrées* combinant les 3 approches précédentes et intégrant très fortement le décideur dans le processus de découverte.

Mesures de Qualité : quantifier l'intérêt d'une règle d'association

Partant d'une structure de données croisant individus et variables binaires (transactions), généralement issue d'un SGBD relationnel, les règles d'associations permettent de découvrir des tendances implicatives entre combinaisons de variables (itemsets). Les premiers travaux sur les règles d'association [4] ont été restreints à l'utilisation de deux mesures d'intérêt aux vertus essentiellement algorithmiques : le support et la confiance. Afin d'améliorer la sélection des meilleures règles, de nombreuses mesures complémentaires ont ensuite été proposées dans la littérature [209]. Freitas [77] distingue les mesures subjectives qui intègrent les buts/connaissances du décideur, et les mesures objectives qui sont des indicateurs statistiques évaluant la contingence d'une règle.

Dans le contexte des mesures objectives, et en nous appuyant sur les travaux précurseurs sur l'analyse statistique implicative de R. Gras [83], et H. Briand et L. Fleury [71], nous proposons d'explorer quatre axes :

1. Le recensement, la classification, et l'étude des propriétés des mesures de qualité disponibles dans la littérature, afin d'inciter à l'usage de mesures autres que le support et la confiance.
2. La conception de nouvelles mesures de qualité issues de l'intensité d'implication, afin de pallier aux limites des mesures recensées.
3. L'étude expérimentale du comportement de ces mesures et de leurs liaisons, afin de faciliter le choix de mesures adaptées aux données traitées et aux besoins du décideur.
4. Le développement d'outils d'étude des mesures et d'aide à la décision pour le choix de "bonnes" mesures.

Fouille de règles : comment faire face au volume

Partant du constat qu'en dépit de la réduction réalisée par les mesures de qualité, le nombre de règles produites demeure prohibitif, nous avons choisi d'intégrer plus fortement le décideur dans le processus d'ECD, lui restituant ainsi son rôle d'acteur majeur intégré au sein même du processus. Ce choix a été doublement motivé, considérant d'une part le nécessaire recentrage sur l'utilisateur prôné par Brachman et Anand [34] ; et d'autre part mes travaux de thèse sur les systèmes anthropocentrés d'aide à la décision [99], menés à l'ENST de Bretagne au sein du Projet JADAR[2] sous la direction de

[2] Jugement, Aide à la Décision et Apprentissage de Règles.

Jean-Pierre Barthélemy. Ce glissement du post-traitement vers une approche interactive d'aide à la décision, où l'utilisateur joue un rôle d'"heuristique" dirigeant la fouille, nous a amené à intégrer trois composantes originales dans notre approche de "fouille anthropocentrée de règles" :

1. Une représentation graphique adaptée aux règles d'association, compatible avec les contraintes cognitives du décideur ;
2. Des opérateurs d'interaction avec la représentation visuelle supportant les besoins de découverte du décideur ;
3. Des algorithmes de fouille locale ad hoc, connectés à la base de données, et pilotés par le décideur à travers les opérateurs d'interaction.

Le principe sous-jacent à cette approche, et issu des contraintes cognitives du décideur, a été dénommé *ciblage de règles* (rule focusing). Il consiste à passer d'une recherche globale (sur la totalité des règles), à une suite de recherches locales (sur un sous-ensemble de règles proches). La progression de t à t+1 dans la suite est réalisée par des opérateurs d'interaction déclenchés sur la représentation visuelle à l'instant t selon les desiderata de l'utilisateur.

Cette approche fortement dynamique, engendre de multiples propriétés sympathiques. La représentation visuelle n'a plus besoin de supporter l'intégralité des règles, ce qui améliore l'intelligibilité des résultats présentés. Les opérateurs incorporent une sémantique locale plus accessible à l'utilisateur et cohérente avec son activité (ex : règles plus générales/spécifiques, règles d'exception,...). Enfin, et cet apport me semble majeur, les algorithmes de fouille ad hoc deviennent locaux, ce qui a le grand avantage de rendre possible la découverte de règles d'association sur des corpus très volumineux (beaucoup de variables), là où les algorithmes globaux comme Apriori, ou les algorithmes sous contraintes, s'effondrent...

Nous proposons de décliner cette approche en deux visualisations complémentaires couplées à des mesures de qualité :

- Par des *graphes,* en considérant l'ensemble des règles comme un immense réseau dont on ne représente qu'une partie à l'aide d'un graphe. La principale difficulté posée réside dans la préservation de la carte mentale de l'utilisateur, risquant d'être perturbée par le caractère fortement dynamique de la représentation, puisque celle-ci doit évoluer partiellement à chaque interaction.

- Par des *métaphores 3D* et l'usage de la *Réalité Virtuelle*, afin d'améliorer l'immersion de l'utilisateur dans la représentation. Les principales difficultés se situent dans la synthèse d'une métaphore intelligible et la définition de voisinages de règles.

Gestion et Ingénierie des connaissances

Parallèlement, stimulés par des projets émanant d'entreprises, et dans une perspective initiale plus applicative, nous nous sommes intéressés à la gestion [68] et l'ingénierie des connaissances [63] et son prolongement récent pour le web sémantique. Le lien avec l'ECD apparaît naturellement lors de la phase de déploiement des connaissances extraites qui est du ressort de la gestion et de l'ingénierie des connaissances. La gestion des connaissances a pour objet le traitement de la connaissance au sens large. Elle concerne l'acquisition, la formalisation, le stockage, la diffusion et la manipulation des connaissances et des savoir-faire généralement détenus par les acteurs (souvent experts) d'une organisation dans un domaine donné. Les deux principales limites rencontrées par les systèmes issus d'une gestion de connaissances concernent leur déploiement dans un système d'information, souvent difficile ; et leurs capacités à évoluer afin de maintenir des connaissances à jour, souvent faibles.

Serveur de connaissances et ontologies

Afin de pallier aux limites de déploiement et d'évolutivité, notre première activité dans le domaine a consisté à concevoir une approche «serveur de connaissances», selon une analogie avec un serveur web transposant ses services à la connaissance et bénéficiant des technologies du web, simplifiant son déploiement. Nous avons implémenté cette approche dans l'outil SAMANTA[3], destiné à la gestion des diagnostics de maintenance, dans le cadre d'un contrat de 5 années avec La Poste. Ce logiciel [182] est fondé sur 3 points de vue ontologiques complémentaires : une ontologie pour les tâches de diagnostics, une ontologie décrivant la composition d'une machine de tri, une ontologie des symptômes. L'outil incorpore un éditeur d'ontologies entièrement graphique garantissant l'évolutivité du système, permet l'association à des ressources multimédia complémentaires (dont des modèles de machine de tri en réalité virtuelle), et enfin un raisonneur prolog permettant d'offrir notamment une fonctionnalité d'aide à la décision pour le diagnostic.

Nous avons ensuite conçu Athanor [74], une généralisation de cette approche serveur de connaissances, incorporant une quatrième ontologie pour modéliser les compétences, et étendant l'ontologie des tâches de diagnostic à des tâches quelconques. Ce logiciel a été réalisé dans le cadre d'un contrat de 3 ans avec la société performanSE SA, et a fait l'objet d'un transfert technologique ayant donné naissance au produit Atanor-knowesia.

[3]Système d'Aide à la MAintenace des Trieuses Automatiques.

Ontologies et Web sémantique

Le Web sémantique [49] s'attache à réintroduire du sens et donc de la connaissance sur le contenu de la Toile. Il offre un ensemble de techniques basées sur XML afin de représenter la sémantique dans des ontologies formelles et d'en permettre le traitement par des requêteurs et des raisonneurs.

Dans cette optique web sémantique, nous nous sommes intéressés à la modélisation de connaissances pour des agents émotionnels supportés par une plateforme multi-agents. A cette fin, nous avons proposé une modélisation agent UML intégrant la dimension émotionnelle dans des agents BDI. Cette formalisation permet de décrire les mécanismes d'évolution interne des agents, ainsi que leurs interactions, et débouche sur la production d'une base de connaissances en RDF/OWL [59], [169]. Ce travail, dont l'objectif in fine est de produire un système d'aide à la décision sur la dynamique des groupes, a été impulsé par le projet ARTA sur l'étude du comportement en milieu professionnel des victimes d'un traumatisme crânien. Il s'est ensuite prolongé dans le cadre d'un projet RIAM : Groupe d'Agents Collaboratifs Emotionnels (GRACE).

Plus récemment, en nous inspirant de nos travaux sur les règles d'association, nous avons proposé une méthode originale, extensionnelle et asymétrique, pour l'alignement d'ontologies définies sous la forme de taxonomies instanciées sur un corpus textuel [53], [54]. L'idée directrice de notre approche est d'offrir un support d'aide à la décision pour l'utilisateur : l'aider à aligner deux ontologies en lui proposant des liens exprimés sous la forme de règles d'association entre concepts. Plus précisément, nous considérons qu'une ontologie est un graphe orienté de concepts structurés par une relation de subsomption, où chaque concept est décrit dans des documents textuels contenant des termes caractéristiques. Notre approche se décompose alors en deux étapes consécutives fondées sur la mesure d'intensité d'implication : (1) l'extraction dans les documents des termes caractéristiques de chaque concept ; puis (2) l'extraction d'un ensemble minimal, au sens d'un critère de réduction des redondances, de règles d'association entre concepts.

Principales contributions

Nos principales contributions peuvent être ventilées en quatre volets : les travaux sur les mesures de qualité, l'approche fouille de règles, les ontologies pour la gestion des connaissances, et le développement d'outils.

Mesures de qualité. Nous avons publié un ensemble de travaux de *synthèse* sur les mesures de qualité, à travers (i) *deux ouvrages collectifs*

[38][105], (ii) un recensement et une *classification* originale de 40 mesures [25, 23], et (iii) une série d'*études comparatives* sur plusieurs base de règles volumineuses, grâce à une approche originale de visualisation des corrélations entre mesures à l'aide *graphes* [120]. Nous avons (iv) développé *trois mesures de qualité* originales : deux mesures entropiques EII [30] et TIC [27], et une mesure statistique IPEE [25]. Nous avons également proposé (v) *plusieurs extensions* de la mesure statistique d'intensité d'implication selon les différents contextes de : données de séquences [17], variables ordinales [95] et floues [86], réduction de variables [48] et élimination des redondances [154], et enfin de cohésion de classes de règles [87].

Fouille de Règles. Nous avons proposé (i) une approche antropocentrée originale pour la *fouille de règles*, basée sur un principe cognitif de fouille locale par *ciblage de règles*, et adossée à une *représentation graphique interactive* munie d'opérateurs de manipulation [152, 16]. Nous avons décliné cette approche en deux représentations interactives originales. D'une part, une représentation par (ii) des *graphes de règles* dotée de capacités dynamiques [146] ; et d'autre part, une représentation par (ii) une *métaphore 3D* représentant des paysages de règles liés par une relation de voisinage [22].

Gestion de connaissances et ontologies. Nous avons conçu (i) une approche *serveur de connaissances* incorporant des modèles ontologiques [73] [103]. Nous avons proposé (ii) une modélisation pour des *agents émotionnels* et leurs interactions, exprimée avec agent UML, puis transposée en rdf/owl. Nous avons conçu (iii) une méthode originale (terminologique, extensionnelle, asymétrique), afin de faciliter l'aide à la décision pour l'*alignement d'ontologie*.

Outils. Les travaux sur les mesures de qualité nous ont amené à implémenter (i) deux outils téléchargeables sur le web : *ARVAL*[4] pour le calcul des mesures, et la plateforme *ARQAT*[5] pour l'étude expérimentale du comportement des mesures. (ii) Deux implémentations de l'approche de fouille de règles ont été réalisées et testées sur des données : *Felix* pour les graphes[155], et *ARVIS*[6] pour les métaphores 3D [22]. Nous avons réalisé (iii) deux implémentations successives de l'approche serveur de connaissances : *SAMANTA*[7] avec La poste [182] [108], et *Athanor* avec PerformanSE SA [74]. (iv) Les agents émotionnels ont été implémentés dans la plateforme multi-agent *TC-plusVirtuel* [59], [60], [169]. Enfin, (v) l'alignement asymétrique d'ontologie a été implémenté dans l'outil *AROMA*[8] [53], [54].

[4]Association Rule VAlidation
[5]Association Rule QuAlity Tool
[6]Association Rule VISualisation
[7]Système d'Aide à la MAiNtenance de Trieuses Automatiques
[8]Association Rule Ontology MAtching

Organisation du document

Ce document est structuré en 5 chapitres selon un progression partant des mesures de qualité pour aller vers la gestion des connaissances. Chaque chapitre ne présente qu'une partie des travaux relatifs à son intitulé ; en contrepartie, chaque chapitre s'achève par un paragraphe "Synthèse des publications" qui récapitule l'ensemble des travaux menés sur la thématique présentée.

Dans un premier temps, nous présentons les résultats de nos travaux sur les mesures de qualité en ECD. Le premier chapitre (**chapitre 2**) introduit le formalisme permettant de définir les notions de règles d'association et de mesure d'intérêt, puis recense à travers une classification originale les mesures d'intérêts repérées dans la littérature. Puis, dans le deuxième suivant (**chapitre 3**) nous définissons trois nouvelles mesures d'intérêt, et en discutons des propriétés et les comparons tant d'un point de vue théorique que sur des expérimentations.

Puis, nous opérons un glissement des mesures de qualité vers l'aide visuelle à la décision, en présentant (**chapitre 4**) une approche instrumentée pour l'évaluation du comportement corrélatif des mesures d'intérêt, que l'on pourrait considérer comme une forme de fouille des mesures d'intérêt. Après une première approche par des techniques de classification de données, nous détaillons les principes des graphes de corrélation. Puis nous proposons une illustration sur une étude comparative expérimentale. Enfin, nous présentons la plateforme ARQAT.

Ce glissement se poursuit vers la fouille de règles d'association dans le chapitre suivant (**chapitre 5**). Nous y présentons une nouvelle approche interactive permettant de réaliser la fouille de règles d'association au sein de représentations graphiques, selon une stratégie de ciblage de règles. Cette approche est ensuite déclinée en une version basée sur des graphes et l'autre sur des métaphores 3D.

Enfin, le dernier chapitre (**chapitre 6**) prolonge ce glissement jusqu'aux connaissances et aux ontologies. Après avoir introduit notre approche "serveur de connaissance" pour la gestion des connaissances, nous présentons les résultats obtenus sur la modélisation de connaissances pour des agents émotionnels, et enfin détaillons une nouvelle approche d'alignement extensionnel et asymétrique d'ontologies.

Chapitre 2

Règles et mesures de qualité

Sommaire

Les mesures de qualité de règles d'association, ou mesures d'intérêt, sont des indicateurs numériques destinés à guider l'utilisateur vers les connaissances potentiellement intéressantes dans les grands volumes de règles produits par les algorithmes de fouille de données. Ces mesures offrent d'évaluer la qualité des règles selon différents points de vue, et permettent d'*ordonner* les règles des meilleures aux plus mauvaises. Elles peuvent également jouer le rôle de *filtre*, en rejetant les règles en dessous d'un seuil de qualité minimale.

Dans les travaux précurseurs sur les règles d'association [4, 5], deux premières mesures fréquentielles sont introduites : le support et la confiance.

Celles-ci sont bien adaptées aux contraintes algorithmiques (cf *apriori*), mais ne sont pas suffisantes pour capturer l'intérêt des règles pour l'utilisateur. Afin de contourner cette limite, de nombreuses mesures d'intérêt complémentaires ont été proposées dans la littérature. Freitas [77, 76] distingue deux types de mesures d'intérêts : les mesures subjectives, et les mesures objectives. Les mesures *subjectives* dépendent des buts, connaissances, croyances de l'utilisateur et sont combinées à des algorithmes supervisés spécifiques afin de comparer les règles extraites avec ce que l'utilisateur connaît ou souhaite [179] [163] [161] [179] [200]. Ainsi, les mesures subjectives proposent de capturer la nouveauté (novelty) ou l'inattendu (unexpectedness) d'une règle par rapport aux connaissances/croyances de l'utilisateur.

En revanche, seuls les cardinaux dus à la contingence des données interviennent dans le calcul des mesures *objectives*. Parmi ces dernières, on trouve aussi bien des mesures fréquentielles rudimentaires que des mesures fondées sur des modèles probabilistes, en passant par des mesures issues de la théorie de l'information ou des mesures statistiques usuelles de liaison [102]. De nombreux travaux de synthèse en récapitulent les définitions et propriétés [11] [114] [208] [209]. Ces synthèses comparent les mesures d'intérêt selon deux aspects différents : d'une part, la définition d'un ensemble de *principes* qui mènent à la conception d'une bonne mesure d'intérêt ; et d'autre part, la *comparaison* des mesures d'intérêt à l'aide de techniques d'analyse de données, afin d'en comprendre le comportement, et in fine d'aider l'utilisateur à choisir les meilleures.

Dans ce chapitre, nous nous intéressons aux mesures objectives de qualité de règles. La subjectivité du post-traitement de règles est prise en compte dans notre approche de fouille de règles grâce aux interactions avec l'utilisateur (voir chapitre 5).

Dans un premier temps nous introduisons la terminologie et les notations qui nous permettent de définir les concepts de règle et de mesure d'intérêt. Puis nous proposons un recensement des principales mesures d'intérêt objectives proposées dans la littérature. Enfin, nous proposons en section 2.3 une classification théorique inédite des mesures de règle. En clarifiant leur signification, une telle classification aide l'utilisateur à quitter l'usage réducteur des seules mesures de confiance et de support, et à étendre son choix vers des indices mieux adaptés à ses besoins.

Ce chapitre s'achève sur une synthèse des publications et des travaux menés sur ce thème.

2.1 Terminologie et notations

2.1.1 Données

Nous considérons un ensemble E de n individus (ou *transactions*) décrits par un ensemble I de p variables booléennes (ou *items*). L'ensemble E est souvent appelée l'*extension* et I l'*intension*. Les p items peuvent être formés par l'expression *variable=modalité,* au moyen d'un codage disjonctif complet appliqué à un ensemble V de variables qualitatives (il peut également s'agir de variables quantitatives discrétisées). En ECD, la relation $E \times I$ est généralement stockée sous forme de table dans une base de données relationnelle.

Un itemset a est un sous-ensemble d'items ($a \subseteq I$) qui représente une conjonction de variables booléennes.

Nous notons A l'extension de a, c'est-à-dire l'ensemble des individus de E qui vérifient a. Formellement, A peut être définie par $E(a) = \{e \in E \,|\, e\, verifie\, a\}$ (correspondance de Galois).

Afin de comptabiliser les effectifs sous-jacents, nous notons $n_a = |A|$ le cardinal de A, et $\mathrm{P}(a) = \frac{n_a}{n}$ la probabilité de a estimée par la fréquence empirique (estimateur du maximum de vraisemblance).

L'ensemble $\overline{A}$, complémentaire de A dans E, est associé à $n_{\overline{a}} = |\overline{A}|$ et $\mathrm{P}(\overline{a}) = \frac{n_{\overline{a}}}{n}$, où $\overline{a}$ est la négation de a.

2.1.2 Règles d'association

Partant de deux itemsets disjoints a et b, nous notons $a \rightarrow b$ la règle d'association traduisant la tendance de b à être vrai quand a est vrai. Elle peut se lire de la manière suivante : "si un individu vérifie a alors il vérifie sûrement b". a est la prémisse de la règle et b sa conclusion.

La table de contingence croisant a et b récapitule les effectifs $n_{ab} = |A \cap B|$, $n_{\overline{a}b} = |\overline{A} \cap B|$, $n_{a\overline{b}} = |A \cap \overline{B}|$, $n_{\overline{a}\overline{b}} = |\overline{A} \cap \overline{B}|$. Le cardinal n_{ab} comptabilise les *exemples* d'une règle (individus de $A \cap B$), c'est-à-dire ceux qui vérifient la prémisse et la conclusion. Tandis que $n_{a\overline{b}}$ comptabilise les contre-exemples (individus de $A \cap \overline{B}$), ceux qui vérifient la prémisse mais pas la conclusion (figure 2.1).

Ainsi, une règle est d'autant meilleure qu'elle admet beaucoup d'exemples et peu de contre-exemples.

On observe les relations suivantes entre les effectifs :

$$n_{ab} + n_{a\overline{b}} = n_a, \quad n_{ab} + n_{\overline{a}b} = n_b, \quad n_{\overline{a}b} + n_{\overline{a}\overline{b}} = n_{\overline{a}},$$

$$n_{a\overline{b}} + n_{\overline{a}\overline{b}} = n_{\overline{b}}, \quad n_a + n_{\overline{a}} = n_b + n_{\overline{b}} = n.$$

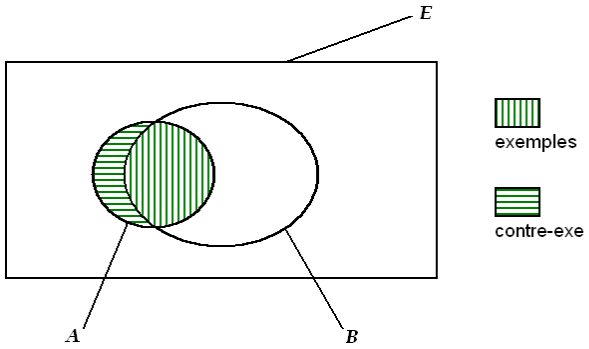


FIG. 2.1 – Diagramme de Venn pour la règle $a \rightarrow b$

La table de contingence croisant a et b ayant 4 degrés de liberté, nous proposons de modéliser les effectifs d'une règle d'association $a \rightarrow b$, sous la forme d'un quadruplet :

$$(n,\, n_a,\, n_b,\, n_{a\bar{b}}).$$

Auquel il faut ajouter les 4 contraintes suivantes :

$$0 \leq n_a \leq n,\ 0 \leq n_b \leq n,\ 0 \leq n_{a\bar{b}} \leq min(n_a,\, n_{\bar{b}}),\ n_{a\bar{b}} \leq n_{\bar{b}},\ \text{et}$$
$$max(0,\, n_a - n_b) \leq n_{a\bar{b}} \leq n_a.$$

2.1.3 Mesures d'intérêt

Reprenant d'une part cette idée qu'une règle est d'autant meilleure que le nombre d'exemples est grand et que le nombre de contre-exemples est petit, et d'autre part en y adjoignant d'autres critères, de nombreuses mesures d'évaluation de la qualité des règles d'association ont été développées. Ces mesures ou indices de qualité sont dénommés *mesures d'intérêt* dans la littérature. Nous proposons de définir une mesure d'intérêt comme une fonction des effectifs d'une règle :

Une ***mesure d'intérêt*** *est une fonction* $I \left| \begin{array}{l} \mathcal{R} \longrightarrow \mathbb{R} \\ (a \rightarrow b) \mapsto I(n, n_a, n_b, n_{a\bar{b}}) \end{array} \right.$.

Les propriétés de la fonction I permettent de distinguer deux catégories de mesures :

- mesures de règles : *I est décroissante avec $n_{a\bar{b}}$ et décroissante avec n_a lorsque les autres variables sont fixes*
- mesures de similarité : *I est positive, symétrique en $n_{a\bar{b}}$ et $n_{\bar{a}b}$, croissante avec n_{ab} et décroissante avec $n_{a\bar{b}}$ lorsque les autres variables sont fixes.*

2.2 Récapitulatif des mesures d'intérêt

2.2.1 Les fondamentaux : Support et confiance

Le support évalue la généralité d'une règle. Il s'agit de la proportion d'individus qui vérifient la règle dans le jeu de données [4] :

$$support(a \rightarrow b) = \frac{n_{ab}}{n}$$

La confiance évalue la validité d'une règle. Il s'agit de la proportion d'individus qui vérifient la conclusion parmi ceux qui vérifient la prémisse [4] :

$$confiance(a \rightarrow b) = \frac{n_{ab}}{n_a}$$

La confiance estime la probabilité conditionnelle que la variable b soit réalisée sachant que la variable a est réalisée. Elle peut aussi être interprétée comme le taux de réussite de la règle.

Le support et la confiance sont des indices simples, et les plus couramment utilisées pour évaluer des règles car ils sont à la base des algorithmes d'extraction de règles (cf Apriori). Toutefois, ils n'offrent qu'une sensibilité limitée aux données, et leurs limites ont été discutées dans nos travaux [96] [97] [98].

2.2.2 Inventaire des mesures

Dans ce chapitre, nous avons recensé les mesures qui sont traditionnellement utilisées comme mesures de règle. Elles sont listées dans le tableau 2.1. Les mesures issues de la théorie de l'information n'y sont pas présentes car elles font l'objet d'une étude spécifique au chapitre 3 (ce sont généralement des mesures de liaison entre variables multimodales et non binaires).

Les sens de variation d'un indice de règle avec n_{ab} et n_a ont été soulignés à l'origine par Piatetsky-Shapiro [183] en tant que propriétés souhaitables d'un indice. Nous les considérons ici comme les fondements de la notion d'indice de règle. Piatetsky-Shapiro considère aussi qu'un bon indice doit décroître avec n_b. Cette condition est trop contraignante pour apparaître dans une définition générale des indices de règle comme la définition 2.1.3, puisque certains indices ne dépendent pas de n_b. Plus généralement, en ce qui concerne les variations par rapport à n_b et n, un indice de règle n'a pas de comportement précis :

- Certains indices ne dépendent pas de n, comme le taux d'exemples et de contre-exemples, la moindre-contradiction, ou l'indice d'Ochiai.

Nom de l'indice *I*	***I*(a→b)=**	**Abrev.**	**Références**
confiance	$\frac{n_{ab}}{n_a}$	Conf	[4]
estimateur laplacien	$\frac{n_{ab}+1}{n_a+2}$	Lapl	[11]
indice de Sebag et Schoenauer	$\frac{n_{ab}}{n_{a\bar{b}}}$	SebSch	[197]
taux des exemples et contre-exemples	$\frac{n_{ab}-n_{a\bar{b}}}{n_{ab}}$	ExRatio	[102]
indice de Ganascia	$\frac{n_{ab}-n_{a\bar{b}}}{n_a}$	Gan	[80]
moindre-contradiction	$\frac{n_{ab}-n_{a\bar{b}}}{n_b}$	LeastC	[8]
indice d'inclusion	$\sqrt[4]{(1-\widehat{H}(b\|a)^2)(1-\widehat{H}(\bar{a}\|\bar{b})^2)}$	IncInd	[91]
indice de Loevinger	$1-\frac{nn_{a\bar{b}}}{n_a n_{\bar{b}}}$	Lov	[166]
coefficient de corrélation	$\frac{nn_{ab}-n_a n_b}{\sqrt{n_a n_b n_{\bar{a}} n_{\bar{b}}}}$	Pears	[180]
rule-interest	$n_{ab}-\frac{n_a n_b}{n}$	RuleInt	[183]
nouveauté	$\frac{n_{ab}}{n}-\frac{n_a n_b}{n^2}$	Nov	[151]
lift ou intérêt	$\frac{nn_{ab}}{n_a n_b}$	Lift	[39]
conviction	$\frac{n_a n_{\bar{b}}}{nn_{a\bar{b}}}$	Conv	[40]
collective strength	$\frac{n_{ab}+n_{\bar{a}\bar{b}}}{n_a n_b+n_{\bar{a}}n_{\bar{b}}}\frac{n^2-n_a n_b-n_{\bar{a}}n_{\bar{b}}}{n-n_{ab}-n_{\bar{a}\bar{b}}}$	ColStren	[2]
indice de Yule	$\frac{n_{ab}n_{\bar{a}\bar{b}}-n_{a\bar{b}}n_{\bar{a}b}}{n_{ab}n_{\bar{a}\bar{b}}+n_{a\bar{b}}n_{\bar{a}b}}$	Yule	[216]
rapport de cotes	$\frac{n_{ab}n_{\bar{a}\bar{b}}}{n_{a\bar{b}}n_{\bar{a}b}}$	OddRat	[173]
multiplicateur de cotes	$\frac{n_{ab}n_{\bar{b}}}{n_{a\bar{b}}n_b}$	OddMul	[150]
Kappa	$\frac{nn_{ab}+nn_{\bar{a}\bar{b}}-n_a n_b-n_{\bar{a}}n_{\bar{b}}}{n^2-n_a n_b-n_{\bar{a}}n_{\bar{b}}}$	κ	[44]
intensité d'implication	$\mathrm{P}(Poisson(\frac{n_a n_{\bar{b}}}{n})>n_{a\bar{b}})$	II	[83]
indice de vraisemblance du lien	$\mathrm{P}(Poisson(\frac{n_a n_b}{n})<n_{ab})$	LL	[159]
opposé de l'indice d'implication	$-\frac{n_{a\bar{b}}-\frac{n_a n_{\bar{b}}}{n}}{\sqrt{\frac{n_a n_{\bar{b}}}{n}}}$	ImpInd	[83]
contribution orientée au χ^2	$\frac{n_{ab}-\frac{n_a n_b}{n}}{\sqrt{\frac{n_a n_b}{n}}}$	ChiCont	[159]
support ou indice de Russel et Rao	$\frac{n_{ab}}{n}$	RusRao	[4] [194]
support causal ou indice de Sokal et Michener	$\frac{n_{ab}+n_{\bar{a}\bar{b}}}{n}$	CauSup	[141] [203]
indice de Rogers et Tanimoto	$\frac{n-n_{a\bar{b}}-n_{\bar{a}b}}{n+n_{a\bar{b}}+n_{\bar{a}b}}$	RogTan	[192]
indice de Jaccard	$\frac{n_{ab}}{n-n_{\bar{a}\bar{b}}}$	Jac	[134]
indice de Dice	$\frac{n_{ab}}{n_{ab}+\frac{1}{2}(n_{a\bar{b}}+n_{\bar{a}b})}$	Dice	[62]
indice d'Ochiai	$\frac{n_{ab}}{\sqrt{n_a n_b}}$	Ochiai	[175]
indice de Kulczynski	$\frac{1}{2}(\frac{n_{ab}}{n_a}+\frac{n_{ab}}{n_b})$	Kulc	[144]

TAB. 2.1 – Les principaux indices de règle

- Certains indices croissent avec n, comme l'indice d'inclusion, l'indice de Yule, l'indice de Rogers et Tanimoto, rule-interest.
- Certains indices décroissent avec n comme le support.
- Certains indices ne sont pas monotones avec n comme la nouveauté.
- Certains indices ne dépendent pas de n_b, comme la confiance, l'indice de Sebag et Schoenauer, le support.
- Certains indices décroissent avec n_b, comme le multiplicateur de cotes, collective strength, l'indice de Jaccard (nous verrons par la suite que tous les indices que nous qualifions d'*écart à l'indépendance* décroissent avec n_b).

2.2.3 Evaluation et sélection des indices

De nombreuses synthèses sur les mesures d'intérêt objectives peuvent être trouvées dans la littérature. Ils traitent principalement deux aspects différents, la définition de l'ensemble de principes d'une bonne mesure d'intérêt, et leur comparaison afin d'aider l'utilisateur à choisir les meilleures.

Plusieurs auteurs se sont intéressés aux propriétés qu'objectivement un bon indice doit vérifier. Ainsi, Piatetsky-Shapiro[183] présente une nouvelle mesure d'intérêt, appelé Rule-Interest, et propose trois principes fondamentaux pour une mesure sur une règle $a \Rightarrow b$: (P1) valeur 0 quand a et b sont indépendants, (P2) croissant avec $a \wedge b$, (P3) décroissant avec a ou b. Dans leur ouvrage, Hiderman et Hamilton [114] proposent cinq principes : minimum value, maximum value, skewness, permutation invariance, transfer. Tan *et al.* [208] [209] définissent cinq principes d'intérêt : symmetry under variable permutation, row/column scaling invariance, anti-symmetry under row/column permutation, inversion invariance, null invariance. Freitas [77] propose un principe de "surprise" d'attribut.

Dans nos travaux [85] [102] [38], nous avons proposé un ensemble de critères, raffinés par [158][150], dont les principaux sont :

- valeur de l'indice quand les variables a et b sont indépendantes,
- valeur de l'indice pour une règle logique,
- variations de l'indice avec n_b,
- variations de l'indice quand tous les effectifs sont dilatés,
- symétrie de l'indice pour une règle et sa réciproque,
- symétrie de l'indice pour une règle et sa règle contraire,
- symétrie de l'indice pour une règle et sa contraposée,
- concavité/convexité de l'indice pour $n_{a\bar{b}} = 0^+$,
- intelligibilité de l'indice,
- facilité à fixer un seuil d'acceptation des règles,
- sensibilité de l'indice aux règles très spécifiques.

D'autres travaux présentent des études comparatives formelles ou expérimentales des indices [11] [207] [209] [212]. Bayardo [11] montre que pour n_b fixé, un certain nombre d'indices sont redondants et les meilleures règles résident le long d'une frontière de support/confiance. Tan *et al.* [207] [209] comparent des mesures symétriques ou symétrisés sur des jeux de règles synthétiques et montrent que les corrélations entre mesures augmente avec la diminution du support. Les indices se révèlent tantôt redondants, tantôt contradictoires, et aucun ne surclasse significativement tous les autres. Kononenco[142] utilise onze mesures pour estimer la qualité des attributs multimodaux, et montre que les valeurs de quatre mesures (information-gain, j-mesure, gini-index, et relevance) tendent à augmenter linéairement avec le nombre de modalités d'un attribut. Zhao et Karypis [217] utilisent huit critères et proposent un algorithme d'optimisation d'un des critères. [81] proposent une classement des mesures par similitude. Vaillant *et al.* [212] évaluent vingt mesures selon 8 critères d'intérêt qui permettent d'identifier trois classes de mesures. Afin de d'assister l'utilisateur dans sa sélection parmi la multitude d'indices candidats, Lenca *et al.* [158] proposent d'appliquer une méthode d'aide multicritère à la décision. Après que l'utilisateur ait exprimé ses préférences sur des critères tels que ceux listés plus haut, la méthode lui fait une recommandation sous la forme d'un ordre partiel sur les indices.

Tous ces travaux s'accordent sur le fait qu'il n'existe pas de mesure idéale. Les mesures se révèlent tantôt redondantes, tantôt contradictoires, et aucune ne surclasse significativement toutes les autres. Le choix des meilleures mesures est fortement contextuel : certaines mesures peuvent être appropriées à certaines applications sur certaines données pour certains utilisateurs, mais pas dans un autre contexte. Une solution pour l'aide au choix des meilleures mesures sera proposée dans le chapitre 4.

2.3 Notre classification des mesures d'intérêt

Dans cette section, nous proposons une classification des mesures d'intérêt selon trois critères : l'*objet* de l'indice, la *portée* de l'indice, et la *nature* de l'indice. Ces critères nous paraissent essentiels pour appréhender la signification des indices, et donc aussi pour aider l'utilisateur à choisir quels indices appliquer.

2.3.1 Objet : indépendance et équilibre

Nous avons vu qu'une règle est d'autant meilleure qu'elle admet beaucoup d'exemples et peu de contre-exemples. Ainsi, pour une règle $a \rightarrow b$ données,

et si l'on ne fait varier que les contre-exemples $n_{a\bar{b}}$ en maintenant les autres effectifs n, n_a, n_b constants ce critère se traduit par : la qualité est minimale lorsque $n_{a\bar{b}} = \min(n_a,\ n_{\bar{b}})$ et maximale lorsque $n_{a\bar{b}} = \max(0,\ n_a - n_b)$.

Entre ces situations, il existe deux configurations spécifiques où l'hypothèse de l'existence d'une règle doit être rejetée : l'*indépendance* et l'*équilibre*. Nous proposons donc de distinguer les mesures selon leur sensibilité aux 3 situations suivantes :

- **Indépendance.** La situation d'indépendance apparaît lorsque les itemsets a et b sont indépendants. C'est-à-dire, lorsque $\mathbf{n_{a\bar{b}} = \frac{n_a n_{\bar{b}}}{n}}$ (équivalent à $\mathrm{P}(a \cap b) = \mathrm{P}(a) \times \mathrm{P}(b)$). Dans ce cas, aucune variable n'apporte aucune information sur l'autre. En conséquence les 8 règles entre a et b ($a \rightarrow b$, $a \rightarrow \bar{b}$, $\bar{a} \rightarrow b$, $\bar{a} \rightarrow \bar{b}$, $b \rightarrow a$, $b \rightarrow \bar{a}$, $\bar{b} \rightarrow a$, et $\bar{b} \rightarrow \bar{a}$) doivent être rejetées.
- **Equilibre.** Nous définissons l'équilibre d'une règle $a \rightarrow b$ comme la situation où la règle possède autant d'exemples que de contre-exemples : $\mathbf{n_{a\bar{b}} = n_{ab} = \frac{1}{2} n_a}$ [26] [23]. Dans cette situation, où l'entropie est maximale, a autant d'incertitude d'amener à b qu'à $\bar{b}$ (maximum d'invraisemblance). Les deux règles $a \rightarrow b$ et $a \rightarrow \bar{b}$ doivent donc être rejetées[1].
- **Aucune.** Certaines mesures ne sont sensibles ni à l'indépendance ni à l'équilibre. Par exemples les mesures de similarité de Rogers et Tanimoto, Jaccard, Dice, Russel et Rao, Skal et Michener.

D'un point de vue général, l'écart à l'équilibre est utile pour prendre des décisions sur b ou faire des prédictions sur b (sachant ou imaginant que a est vrai, b est-il vrai ou faux ?), tandis que l'écart à l'indépendance est utile pour découvrir des liaisons entre a et b (la vérité de a influence-t-elle la vérité de b ?).

L'écart à l'équilibre et l'écart à l'indépendance doivent être considérés comme complémentaires. En particulier, il ne faut pas négliger les indices d'écart à l'équilibre au profit des indices d'écart à l'indépendance quand les réalisations des variables étudiées sont rares. De nombreux auteurs citent pourtant parmi les propriétés majeures d'un bon indice de règle le principe suivant (énoncé à l'origine dans [183]) : "un indice doit s'annuler (ou prendre une valeur fixe) à l'indépendance" [209] [85] [150] [158]. Ce principe nie totalement la notion d'écart à l'équilibre ; l'utiliser revient à considérer que les indices d'écart à l'indépendance mesurent mieux la qualité des règles que les indices d'écart à l'équilibre.

D'autre part l'indépendance se définit à l'aide des quatre paramètres n_{ab}, n_a, n_b, et n, alors que l'équilibre n'en nécessite que deux : n_{ab} et n_a. Les indices

[1] Les mesures d'intérêt sensibles à ces deux situations devraient non seulement permettre leur rejet, mais aussi assurer un intervalle de sécurité à leur voisinage .

d'écart à l'indépendance sont généralement décroissant avec n_b. Les seules exceptions à ce principe sont l'indice d'inclusion et la moindre-contradiction.

2.3.2 Portée : quasi-implication, quasi-conjonction et quasi-équivalence

De prime abord, on peut penser qu'une règle d'association est une tendance à l'implication logique qui admettrait des contre-exemples, ce qu'on appelle communément une "quasi-implication". Cependant, dans le cadre logique, une implication $a \rightarrow b$ est équivalente à sa contraposée $\overline{b} \rightarrow \overline{a}$. Or, cette propriété n'est plus vérifiée sur les règles d'association. Plus précisément, derrière la notation $a \rightarrow b$, Kodratoff fait la distinction entre deux types de liaisons de sémantique implicative pouvant être découvertes dans des données [141], la *co-occurrence* et la *causalité,* et établit le lien avec le paradoxe de Hempel[2] lié à la prise en compte de la contraposée,

Ainsi, nous proposons de distinguer la portée d'une règle d'association $a \rightarrow b$ à travers 4 situations différentes :

1. La *règle d'association* de référence $a \rightarrow b$, dont les contre-exemples sont mesurés par $n_{a\overline{b}}$ et les exemples par n_{ab}.
2. La *quasi-implication* $a \Rightarrow b$, où l'on considère à la fois la règle $a \rightarrow b$ et sa contraposée $\overline{b} \rightarrow \overline{a}$, dont les contre-exemples sont mesurés par $n_{a\overline{b}}$ et les exemples par n_{ab} et $n_{\overline{a}\overline{b}}$. $a \Rightarrow b$ *est donc équivalente à sa contraposée* $\overline{b} \Rightarrow \overline{a}$.
3. La *quasi-conjonction* $a \leftrightarrow b$, où l'on considère à la fois la règle $a \rightarrow b$ et sa symétrique $b \rightarrow a$, dont les contre-exemples sont mesurés par $n_{a\overline{b}}$et $n_{\overline{a}b}$ et les exemples par n_{ab}. $a \leftrightarrow b$ *est donc équivalente à sa symétrique* $b \leftrightarrow a$.
4. La *quasi-équivalence* $a \Leftrightarrow b$, où l'on considère à la fois la quasi-implication $a \Rightarrow b$ et sa symétrique $b \Rightarrow a$, dont les contre-exemples sont mesurés par $n_{a\overline{b}}$et $n_{\overline{a}b}$ et les exemples par n_{ab} et $n_{\overline{a}\overline{b}}$. $a \Leftrightarrow b$ repose donc sur les quatre règles $a \rightarrow b$, $b \rightarrow a$, $\overline{b} \rightarrow \overline{a}$, et $\overline{a} \rightarrow \overline{b}$. $a \Leftrightarrow b$ *est équivalente à sa symétrique* $b \Leftrightarrow a$.

[2]Le paradoxe de Hempel réside dans le fait qu'un énoncé comme "Tous les corbeaux sont noirs" (logiquement équivalent à "Tout ce qui n'est pas noir n'est pas corbeau") est confirmé par l'observation d'une chaise blanche, d'une chaussure marron, d'un courant d'air...

2.3.3 Nature : descriptive versus statistique

Notre troisième et dernier critère de classification est la nature *descriptive* ou *statistique* des mesures de règle. Il est aussi recensé dans [102],[150] et dans [85].

- **Mesures descriptives.** Les mesures descriptives (ou fréquentielles) ne varient pas avec la dilatation des effectifs (quand tous les effectifs des données sont augmentés ou diminués selon la même proportion). Elles vérifient $I(n_{ab},\ n_a,\ n_b,\ n) = I(\alpha.n_{ab},\ \alpha.n_a,\ \alpha.n_b,\ \alpha.n)$ pour toute constante α strictement positive. Ces mesures prennent en compte les tailles des ensembles d'individus A, B, et $A \cap B$ uniquement de manière relative (par les probabilités $\mathrm{P}(a)$, $\mathrm{P}(b)$, $\mathrm{P}(a \cap b)$) et non de manière absolue (par les effectifs n_a, n_b, n_{ab}).
  Les mesures de règle issus de la théorie de l'information (mesures entropiques) sont toutes de nature descriptives. Ils sont étudiés au chapitre 3.
- **Mesures statistiques.** Les mesures statistiques sont celles qui varient avec la dilatation des effectifs. Elles tiennent compte de la taille des phénomènes étudiés, c'est-à-dire que les effectifs n, n_a, n_b, n_{ab} sont considérés de manière absolue. Statistiquement, une règle est en effet d'autant plus fiable qu'elle est évaluée sur un grand volume de données. Parmi les mesures de règle de nature statistique, il existe une catégorie spécifique : celle des mesures probabilistes. Ces mesures calculent généralement la probabilité de vérifier une hypothèse statistique (1 - l'erreur de première espèce ou p-value d'un test d'hypothèse sur une variable aléatoire). Parmi celles-ci citons : le χ^2 [195], l'intensité d'implication [83], et l'indice de vraisemblance du lien [159].

2.3.4 Classification en Objet, Nature et Portée

La classification des mesures de règle selon l'objet, la portée, et la nature est donnée dans le tableau 2.2. Certaines cellules du tableau sont vides. Tout d'abord, une mesure de similarité étant symétrique par permutation des variables, il ne peut s'agir ni d'une mesure de règle au sens strict, ni d'une mesure de quasi-implication au sens strict. Ensuite, il n'existe aucune mesure de quasi-conjonction ou de quasi-équivalence qui soit sensible à l'écart à l'équilibre. De telles mesures pourraient être développées, mais ils nécessiteraient d'associer des règles dont les équilibres sont différents. Contrairement à l'indépendance, l'équilibre d'une règle $a \rightarrow b$ n'est en effet ni l'équilibre de $b \rightarrow a$, ni celui de $\overline{b} \rightarrow \overline{a}$, ni celui de $\overline{a} \rightarrow \overline{b}$. La seule mesure qui combine des équilibres différents (en l'occurrence ceux d'une règle et de sa contraposée)

est l'indice d'inclusion.

Alors que l'on considère généralement que les mesures de règle sont très nombreuses, pour ne pas dire trop, la classification montre qu'il existe en fait peu de mesures de règle au sens strict. En particulier, la seule mesure d'écart à l'indépendance qui porte sur des règles au sens strict est le multiplicateur de cotes [150]. Par ailleurs, il n'existe aucune mesure statistique sensible à l'écart à l'équilibre. Nous en proposons une dans le chapitre suivant (chapitre 3).

2.4 Synthèse des publications

Nos premiers travaux effectués sur les mesures de qualité en ECD, ont consisté à diagnostiquer les limites de la confiance et du support, puis à expliquer l'intérêt d'utiliser l'intensité d'implication afin de contourner certaines de ces limites [96] [97] [98].

Dans le cadre de l'animation du groupe de travail national GafoQualité[3] [100], nous avons réalisé un ensemble de travaux de synthèse sur les mesures de qualité [101] [84] [104]. Cette synthèse nationale a abouti à la publication d'un numéro spécial, intitulé *mesures de qualité Mesures de qualité pour la fouille de données*, dans revue *RNTI* [38] en 2004, qui s'est prolongé par l'édition d'un ouvrage collectif international, intitulé *Quality Measures in Data Mining*, publié par Springer, dans la série *Studies in Computational Intelligence* [105]. Les activités de ce groupe se poursuivent aujourd'hui à travers la tenue d'un atelier annuel *Qualité des Données et des Connaissances*[4] [13] [14]. Notons également l'impulsion donnée à l'échelle nationale à travers l'organisation de la première conférence EGC [37].

Nos travaux ont été valorisés à un niveau international au cours d'une session invitée [107] à la conférence ASMDA en 2005, et au niveau national lors d'un tutoriel sur les mesures de qualité [102], où j'ai proposé un récensement et une synthèse sur ce sujet.

Un recensement des mesures de qualité et de leurs propriétés a été pousuivi dans le cadre de deux travaux de thèse [16], [119].

Enfin, nous avons proposé une classification originale des mesures de qualité recensées dans la littérature selon 3 critères [25] [23] [24] [27] [16]. Cette classification théorique nous a permis de mettre en évidence l'absence de me-

[3]Le groupe de travail GafoQualité a été animé par R. Gras et F. Guillet en 2002 et 2003, au sein de l'AS GafoDonnées STIC CNRS. Il a rassemblé pendant 2 ans une trentaine de chercheurs répartis sur une dizaine de laboratoires nationaux.

[4]atelier QDC dans le cadre des conférences EGC 2005 et 2006, co-organisé par L. Berti-Equille et F. Guillet.

Portée /Objet	Ecart à l'équilibre	Ecart à l'indépendance	Similarité
Règle	– confiance, – indice de Sebag et Schoenauer, – taux des exemples et contre-exemples, – estimateur laplacien de probabilité conditionnelle, – indice de Ganascia, – moindre-contradiction	– multiplicateur de cotes	
Quasi-implication	– indice d'inclusion	– indice de Loevinger, – conviction, – *intensité d'implication*, – *indice d'implication*	
Quasi-conjonction		– lift ou intérêt – *indice de vraisemblance du lien*, – *contribution orientée au* χ^2	– support ou indice de Russel et Rao, – indice de Jaccard, – indice de Dice, – indice d'Ochiai, – indice de Kulczynski
Quasi-équivalence		– coefficient de corrélation, – nouveauté, – collective strength, – κ, – indice de Yule, – rapport de cotes – *rule-interest*	– support causal ou indice de Sokal et Michener, – indice de Rogers et Tanimoto

NB : La **nature** est indiquée par le style de la police : les mesures en *italique* sont statistiques, les autres sont descriptives.

TAB. 2.2 – Classification des indices de règle

sures sur certaines combinaisons de critères, et nous a amené à définir de nouvelles mesures pour y répondre (détaillées au chapitre suivant).

Chapitre 3

Conception de Nouvelles Mesures

Sommaire

Dans ce chapitre, nous présentons trois nouvelles mesures de règle issues de nos travaux : l'*indice probabiliste d'écart à l'équilibre*, l'*intensité d'implication entropique*, et le *taux informationnel modulé par la contraposée*. La première mesure, nommé IPEE, est la seule mesure d'écart à l'équilibre qui soit de nature statistique. Elle est fondée sur un modèle probabiliste et évalue

la significativité de l'écart à l'équilibre. L'intensité d'implication entropique (IIE), quant à elle, est une extension de l'intensité d'implication hybridée par un indice entropique d'inclusion, afin de l'adapter aux données volumineuses. Elle prend en compte à la fois l'écart à l'équilibre et l'écart à l'indépendance. Enfin, le taux informationnel modulé par la contraposée (TIC) est une mesure fondée sur la théorie de l'information. Elle possède la particularité unique de rejeter simultanément les mauvais écarts à l'équilibre et à l'indépendance. Pour chacune des ces trois mesures de règle, nous décrivons sa construction et étudions ses propriétés sur des simulations numériques.

Ce chapitre s'achève sur une synthèse des publications et des travaux menés sur ce thème.

3.1 Fondements : l'Intensité d'Implication

L'*intensité d'implication* est une mesure d'intérêt proposée par Régis Gras [83] dans le cadre de *l'analyse statistique implicative* [90]. Cette mesure, initialement appliquée à la didatique des mathématique, a ensuite été utilisée en ECD sous l'impulsion d'Henri Briand et Laurent Fleury [71]. Cette mesure quantifie l'invraisemblance de la faiblesse du nombre de contre-exemples $n_{a\bar{b}}$ relativement à une l'hypothèse d'indépendance entre a et b.

Pour une règle $a \rightarrow b$, l'intensité d'implication compare le nombre de contre-exemples $n_{a\bar{b}}$ observé dans les données au nombre de contre-exemples attendu sous l'hypothèse H_0 d'indépendance entre a et b. Associons donc aux ensembles A et B deux ensembles indépendants X et Y tirés aléatoirement dans E et de mêmes cardinaux que A et B : $|X| = n_a$ et $|Y| = n_b$ (figure 3.1). Le nombre de contre-exemples attendu sous H_0 est le cardinal $|X \cap \overline{Y}|$. Il s'agit d'une variable aléatoire, notée $\mathcal{N}_{a\bar{b}}$, dont $n_{a\bar{b}}$ est une valeur observée. La règle $a \rightarrow b$ est d'autant meilleure que la probabilité que le hasard produise plus de contre-exemples que les données est grande.

L'*intensité d'implication* φ d'une règle $a \rightarrow b$ est définie par :

$$\varphi(a \rightarrow b) = \mathrm{P}(\mathcal{N}_{a\bar{b}} > n_{a\bar{b}} \ \mid \mathrm{H}_0)$$

La variable aléatoire $\mathcal{N}_{a\bar{b}}$ peut être modélisée selon différentes lois :

- loi *hypergéométrique* de paramètres $(n, n_a, n_{\bar{b}})$: $P(\mathcal{N}_{a\bar{b}} = k) = \frac{C_{n_{\bar{a}}}^{n_{\bar{b}}-k} C_{n_a}^{k}}{C_n^{n_b}}$
- loi *binomiale* de paramètres $(n, \frac{n_a n_{\bar{b}}}{n^2})$: $P(\mathcal{N}_{a\bar{b}} = k) = C_{n_a}^{k} \left(\frac{n_{\bar{b}}}{n}\right)^k . \left(\frac{n_b}{n}\right)^{n_a - k}$
- loi de *Poisson* de paramètre $\frac{n_a n_{\bar{b}}}{n}$: $P(\mathcal{N}_{a\bar{b}} = k) = \frac{\lambda^k}{k!} e^{-\lambda}$, avec $\lambda = \frac{n_a n_{\bar{b}}}{n}$ (nombre de contre-exemples à l'indépendance)

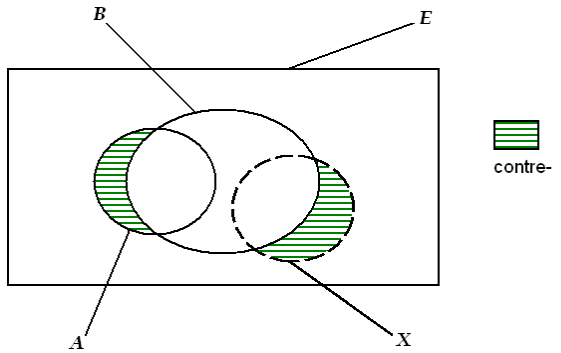


FIG. 3.1 – Modèle aléatoire pour les contre-exemples $\mathcal{N}_{a\bar{b}}$

Obtient alors φ par intégration de la loi choisie :

$$\varphi(a \rightarrow b) = 1 - \sum_{k=max(0,n_a-n_b)}^{min(n_{a\bar{b}},n_{\bar{b}})} P(\mathcal{N}_{a\bar{b}}=k)$$

De plus, quand $\lambda > 15$, la loi de Poisson peut être approximée par une loi normale, et l'intensité d'implication s'écrit alors :

$$\varphi(a \rightarrow b) = \mathrm{P}(\mathcal{N}(0,1) > \widetilde{n}_{a\bar{b}}) = \frac{1}{\sqrt{2\pi}} \int_{\widetilde{n}_{a\bar{b}}}^{\infty} e^{-\frac{t^2}{2}} \mathrm{dt}$$

Où $\widetilde{n}_{a\bar{b}} = \frac{n_{a\bar{b}}-\lambda}{\sqrt{\lambda}}$ est la variable $n_{a\bar{b}}$ centrée réduite. $\widetilde{n}_{a\bar{b}}$ est appelé indice d'implication [83].

3.2 L'intensité d'implication entropique (*IIE*)

Comme toutes les mesures statistiques probabilistes, cette mesure devient malheureusement peu discriminante lorsque les cardinaux étudiés deviennent grands, puisque la variance de la loi mesurée tend vers 0 quand n croit. Pour résoudre ce problème, Nous avons proposé dans Gras *et al.* [91] de moduler les valeurs de l'intensité d'implication par un indice descriptif de quasi-implication fondé sur l'entropie de Shannon : l'*indice d'inclusion*. La nouvelle mesure obtenue a été nommée *intensité d'implication entropique* (IIE). Elle a l'originalité de prendre en compte à la fois l'écart à l'équilibre et l'écart à l'indépendance.

Ce qui nous amène à définir auparavant l'indice d'inclusion et l'entropie réduite.

3.2.1 L'Entropie réduite

Rappelons que l'entropie d'une variable binaire a est définie par :

$$H(a) = H(P(a), P(\overline{a})) = -P(a).\log_2 P(a) - P(\overline{a}).\log_2 P(\overline{a})$$

Ce qui nous donne les expressions de $H(a)$, et de $H(a|b)$ l'entropie de a conditionnelle à b :

$$H(a) = -\frac{n_a}{n}.\log_2 \frac{n_a}{n} - \frac{n_{\overline{a}}}{n}.\log_2 \frac{n_{\overline{a}}}{n} \text{ , et } H(a/b) = -\frac{n_{ab}}{n_b}.\log_2 \frac{n_{ab}}{n_b} - \frac{n_{\overline{a}b}}{n_b}.\log_2 \frac{n_{\overline{a}b}}{n_b}.$$

Afin de supprimer la symétrie introduite par l'entropie, nous proposons d'utiliser une fonction entropique orientée $\widehat{H}$ appelée *entropie réduite* [26]. L'*entropie réduite* $\widehat{H}(a)$ d'une variable binaire a est définie par : $\widehat{H}(a) = H(a)$ si $P(a) \leq \frac{1}{2}$ (équivalent $n_a \leq \frac{n}{2}$), et $\widehat{H}(a) = 1$ sinon. On définit de la même manière l'*entropie conditionnelle réduite* d'une variable binaire b sachant la réalisation de a, par :

$$\widehat{H}(b/a) = H(b/a) \text{ si } P(b/a) \leq \tfrac{1}{2} \text{ (équivalent } n_{ab} \leq \tfrac{n_a}{2} \text{),}$$

$$\text{et } \widehat{H}(b/a) = 1 \text{ sinon.}$$

Notons que l'entropie $H(a)$ peut alors s'écrire comme la somme de deux entropies réduites : $H(a) = \widehat{H}(a) + \widehat{H}(\overline{a}) - 1$.

3.2.2 L'indice d'inclusion

L'indice d'inclusion est une mesure *descriptive entropique* destinée à mesurer l'*écart à l'équilibre* de la *quasi-implication* $a \Rightarrow b$, c'est-à-dire du couple $a \rightarrow b$ et $\overline{b} \rightarrow \overline{a}$ (sa contraposée). Comme nous l'avons vu dans le chapitre 2, l'équilibre d'une règle n'est pas l'équilibre de sa contraposée (contrairement à l'indépendance qui, elle, est commune aux deux règles). L'idée fondatrice est de mesurer l'entropie $H(a \rightarrow b)$ de la règle $a \rightarrow b$ par son entropie conditionnelle $H(b|a)$. Toutefois, comme cette entropie est symétrique en $P(b|a) = \frac{1}{2}$, nous avons choisi d'utiliser sa forme asymétrique, l'*entropie réduite* $\widehat{H}(b/a)$.

L'*indice d'inclusion* $\tau(a \rightarrow b)$ est alors définit comme la combinaison des entropies réduites des règles $a \rightarrow b$ et $\overline{b} \rightarrow \overline{a}$:

$$\tau(a \rightarrow b) = \sqrt[2\omega]{(1 - \hat{H}^{\omega}(a \rightarrow b))(1 - \hat{H}^{\omega}(\overline{b} \rightarrow \overline{a}))}$$

où ωest un réel strictement positif

Cet indice d'inclusion présente l'originalité d'être la seule mesure d'écart à l'équilibre qui porte sur des quasi-implications (voir classification 2.2 page 21). Il s'annule dès que l'écart à l'équilibre de la règle ou de sa contraposée n'est pas orienté en faveur des exemples, c'est-à-dire lorsque $n_{a\bar{b}} \geq \min(\frac{n_a}{2}, \frac{n_{\bar{b}}}{2})$.

ω est un paramètre de sélectivité de l'indice d'inclusion qui peut être ajusté en fonction des données étudiées : plus ω est faible, plus l'indice d'inclusion décroît rapidement avec les contre-exemples, et plus le filtrage des règles est sévère (cf [17]). En analyse statistique implicative, c'est généralement $\omega = 2$ qui est retenu. Ce choix engendre un indice d'inclusion qui réagit faiblement aux premiers contre-exemples, ce qui est pour nous une propriété fondamentale d'un bon indice de règle [85].

3.2.3 L'intensité d'implication entropique

L'*intensité d'implication entropique* ϕ *d'une règle* $a \rightarrow b$ est alors définie comme la moyenne géométrique de l'intensité d'implication $\varphi(a \rightarrow b)$ et de l'indice d'inclusion $\tau(a \rightarrow b)$ [91] :

$$\phi(a \rightarrow b) = \sqrt{\varphi(a \rightarrow b) \times \tau(a \rightarrow b)}$$

3.3 L'Indice Probabiliste d'Ecart à l'Equilibre (*IPEE*)

Nous avons vu au chapitre précédent qu'il n'existe aucune mesure *statistique* sensible à l'*écart à l'équilibre*. Dans [23, 25], nous avons donc proposé de combler cette lacune en définissant une nouvelle mesure : *l'indice probabiliste d'écart à l'équilibre (IPEE).* Cette mesure est fondée sur un modèle probabiliste inspiré de l'intensité d'implication. Plus précisément, cette mesure évalue la significativité de l'écart à l'équilibre, là où l'intensité d'implication ou l'indice de vraisemblance du lien évaluent la significativité de l'écart à l'indépendance.

Etant donnée une règle $a \rightarrow b$, nous cherchons à mesurer la significativité statistique de l'écart à l'équilibre de la règle. La configuration d'équilibre nous amène à définir une hypothèse de référence H_0 correspondant à l'équiprobabilité entre les exemples et les contre-exemples ($n_{a\bar{b}} = n_{ab}$). Associons donc à l'ensemble $A = E(a)$ un ensemble aléatoire X de même cardinal n_a tiré dans

E sous cette hypothèse H_0, où : $\mathrm{P}(X \cap B) = \mathrm{P}(X \cap \overline{B})$. Le nombre de contre-exemples est $\mathcal{N}_{a\bar{b}}$ attendu sous H_0 est le cardinal de $X \cap \overline{B}$ ($\mathcal{N}_{a\bar{b}}$=$|X \cap \overline{B}|$), et l'effectif $n_{a\bar{b}}$ est le nombre de contre-exemples observé sur la règle $a \rightarrow b$.

L'*indice probabiliste d'écart à l'équilibre (IPEE)* d'une règle $a \rightarrow b$ est alors défini par :

$$IPEE(a \rightarrow b) = \mathrm{P}(\mathcal{N}_{a\bar{b}} > n_{a\bar{b}} \mid \mathrm{H}_0)$$

Dans le cadre d'un tirage avec remise, $\mathcal{N}_{a\bar{b}}$ suit une loi binomiale de paramètres $\frac{1}{2}$ (autant de chances de tirer un exemple que de tirer un contre-exemple) et n_a. IPEE s'écrit donc :

$$IPEE(a \rightarrow b) = 1 - \frac{1}{2^{n_a}} \sum_{k=0}^{n_{a\bar{b}}} \mathrm{C}_{n_a}^k$$

Quand $n_a > 15$, la loi binomiale peut être approximée par une loi normale, et IPEE s'écrit alors :

$$IPEE(a \rightarrow b) = \mathrm{P}(\mathcal{N}(0,1) > \widetilde{n}_{a\bar{b}}) = \frac{1}{\sqrt{2\pi}} \int_{\widetilde{n}_{a\bar{b}}}^{\infty} e^{-\frac{t^2}{2}} \mathrm{dt}$$

où $\mathcal{N}(0,1)$ est la loi normale centrée réduite, et $\widetilde{n}_{a\bar{b}} = \frac{n_{a\bar{b}} - \frac{n_a}{2}}{\sqrt{\frac{n_a}{4}}}$ la valeur observée centrée et réduite.

IPEE quantifie donc l'invraisemblance de la petitesse du nombre de contre-exemples $n_{a\bar{b}}$ relativement à l'hypothèse H_0.

En particulier :

- si $IPEE(a \rightarrow b)$ vaut 0 (resp. 1), alors il est invraisemblable que les caractères (a *et* b) et (a *et* $\bar{b}$) soient équiprobables (l'écart à l'équilibre de la règle est significatif mais orienté en faveur des contre-exemples (resp. des exemples)) ;
- si $IPEE(a \rightarrow b)$ vaut 0.5, alors il est vraisemblable que les caractères (a *et* b) et (a *et* $\bar{b}$) soient équiprobables (l'écart à l'équilibre de la règle n'est pas significatif) ;

De manière analogue à l'intensité d'implication, cette nouvelle mesure peut être interprétée comme le complément à 1 de la probabilité critique (*p-value*) d'un test d'hypothèse.

3.4 Le taux informationnel modulé par la Contraposée (*TIC*)

Parmi les mesures utilisées en ECD pour évaluer la qualité des règles, celles issues de la théorie de l'information [198] sont particulièrement intelligibles et utiles puisqu'elles peuvent être interprétées en termes d'apport d'information. Les mesures d'intérêt issues de la théorie de l'information sont l'entropie conditionnelle de Shannon [43], l'information mutuelle moyenne [135], le coefficient d'incertitude de Theil [210] [191] [209], la J-mesure [202], et l'indice de Gini [11] [135]. Ces mesures entropiques ne sont cependant pas des mesures de règle au sens de la définition 2.1.3. Il s'agit en effet de mesures de liaison entre variables multimodales, qui traitent identiquement exemples et contre-exemples. En particulier, elles ne permettent pas de distinguer les règles contraires $a \rightarrow b$ et $a \rightarrow \overline{b}$ alors même qu'elles ont des significations opposées. Ces mesures sont davantage adaptées à l'évaluation de règles de classification en apprentissage supervisé, où le modèle recherché doit expliquer toutes les modalités de la variable classe.

Partant de l'*indice d'inclusion* [91] (cf. section 3.2.2 page 26) qui est une mesure d'*écart à l'équilibre*, nous avons proposé de développer une nouvelle mesure entropique qui soit également sensible à l'*écart à l'indépendance*. Cette nouvelle mesure, appelée *taux informationnel*, a l'originalité d'être simultanément sensible aux *écarts à l'équilibre* et *à l'indépendance*.

3.4.1 Taux informationnel *(TI)*

L'idée est d'orienter le gain d'information (ou taux d'information) $g(a \rightarrow b) = \frac{H(b)-H(b/a)}{H(b)}$ en le rendant asymétrique grâce à l'utilisation de l'entropie réduite $\hat{H}$ (cf. section 3.2.1). Ainsi, le *Taux Informationnel* TI apporté par a dans la règle $a \rightarrow b$ est défini par :

$$TI(a \rightarrow b) = \frac{\widehat{H}(b)-\widehat{H}(b/a=1)}{\widehat{H}(b)} \text{ , si } n_{\overline{b}} \neq 0$$

3.4.2 Taux Informationnel modulé par la Contraposée *(TIC)*

En associant les taux informationnels d'une règle $a \rightarrow b$ et de sa contraposée $\overline{b} \rightarrow \overline{a}$ au sein d'une mesure synthétique, l'indice de règle TI peut être dérivé en une mesure de quasi-implication. Afin que le contenu informationnel de la quasi-implication soit nul dès que la règle ou sa contraposée n'est

pas informative, nous avons choisi d'utiliser la moyenne géométrique pour combiner les deux taux informationnels en rejetant tous les taux négatifs.

Le *Taux Informationnel modulé par la Contraposée (TIC)* d'une règle $a \rightarrow b$ est défini par :

$$TIC(a \rightarrow b) = \sqrt{TI(a \rightarrow b) \times TI(\overline{b} \rightarrow \overline{a})}, \text{ si } TI(a \rightarrow b) \geq 0 \text{ et } TI(\overline{b} \rightarrow \overline{a}) \geq 0$$

$$TIC(a \rightarrow b) = 0, \text{ sinon}$$

3.5 Propriétés

3.5.1 Propriétés d'II

Objet	écart à l'indépendance
Portée	quasi-implication
Nature	statistique
Domaine de variation	[0 ; 1]
Valeur pour les règles logiques	$1 - e^{-\frac{n_a n_{\overline{b}}}{n}}$
Valeur pour les règles à l'équilibre	< 1
Valeur pour les règles à l'indépendance	0

TAB. 3.1 – Propriétés de l'intensité d'implication

Les principales propriétés de l'intensité d'implication sont résumées dans le tableau 3.1. Dans la figure 3.2, l'indice est représenté en fonction du nombre de contre-exemples et comparé aux principaux indices d'écarts à l'indépendance : l'indice de Loevinger, le lift, et la corrélation. Nous pouvons voir que :

- L'intensité d'implication réagit faiblement aux premiers contre-exemples (décroissance lente). Ce comportement est intuitivement satisfaisant pour un indice statistique puisqu'un faible nombre de contre-exemples ne saurait remettre en cause la règle [85].
- Le rejet des règles s'accélère dans une zone d'incertitude autour de l'indépendance $n_{a\overline{b}} = \frac{n_a n_{\overline{b}}}{n}$ (décroissance rapide).

Dans les figures 3.3, les effectifs des données sont multipliés par un coefficient γ à partir d'une configuration initiale. Les indices sont représentés en fonction de γ. Seule l'intensité d'implication est de nature statistique et prend

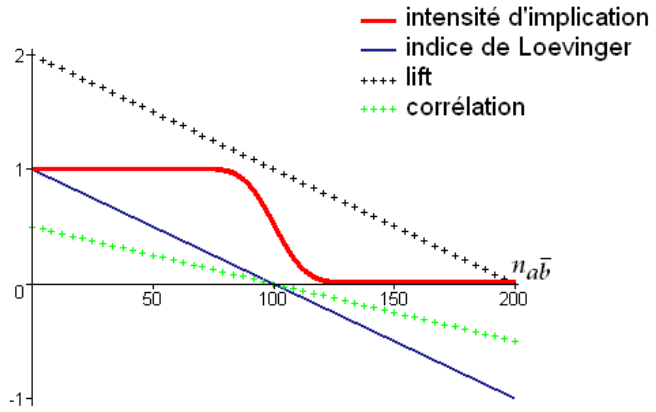
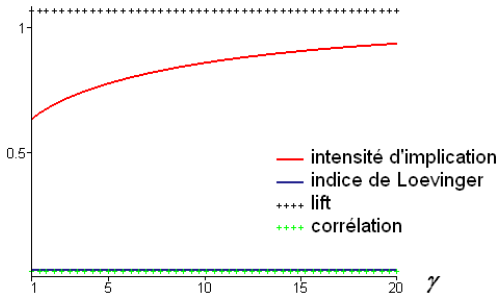
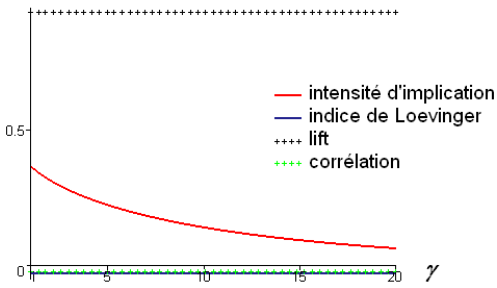


FIG. 3.2 – Représentation des indices d'écart à l'indépendance selon $n_{a\bar{b}}$ ($n_a = 200,\ n_b = 500,\ n = 1000,\ n_{a\bar{b}} \in [0\ ;\ 200]$)



(A) $n_a = 200 \times \gamma,\ n_b = 300 \times \gamma,$ $n_{a\bar{b}} = 136 \times \gamma,\ n = 1000 \times \gamma,$ $\gamma \in [1\ ,\ 20]$



(B) $n_a = 200 \times \gamma,\ n_b = 300 \times \gamma,$ $n_{a\bar{b}} = 144 \times \gamma,\ n = 1000 \times \gamma,$ $\gamma \in [1\ ,\ 20]$

FIG. 3.3 – Représentation des indices d'écarts à l'indépendance en fonction de la dilatation des effectifs

en compte les cardinaux des données de manière absolue : plus le coefficient multiplicateur γ est grand, plus l'écart à l'indépendance observé dans les données est statistiquement significatif, et plus on peut confirmer (figure 3.3.(A)) ou infirmer (figure 3.3.(B)) la bonne qualité de la règle. Cependant, comme le montre la figure 3.4, l'intensité d'implication devient peu discriminante quand les cardinaux étudiés sont grands.

3.5.2 Propriétés d'IIE

L'intensité d'implication entropique prend en compte à la fois l'écart à l'indépendance et l'écart à l'équilibre. Toutefois, selon la classification du chapitre 2, l'intensité d'implication entropique est uniquement un indice d'écart

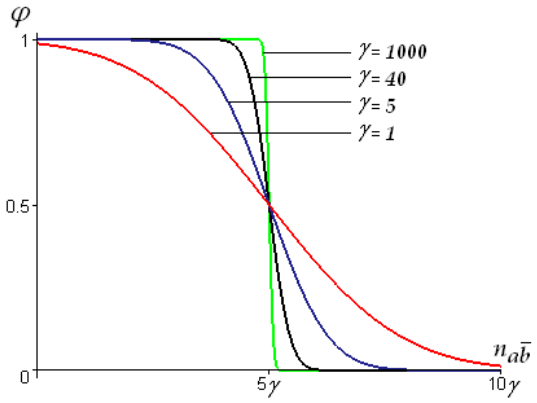


FIG. 3.4 – Représentation de l'intensité d'implication avec la dilatation des effectifs
($n_a = 20 \times \gamma,\ n_b = 75 \times \gamma,\ n = 100 \times \gamma,$
$n_{a\bar{b}} \in [0 \times \gamma\ ;\ 10 \times \gamma],\ \gamma \in \{1;\ 5;\ 40;\ 1000\}$)

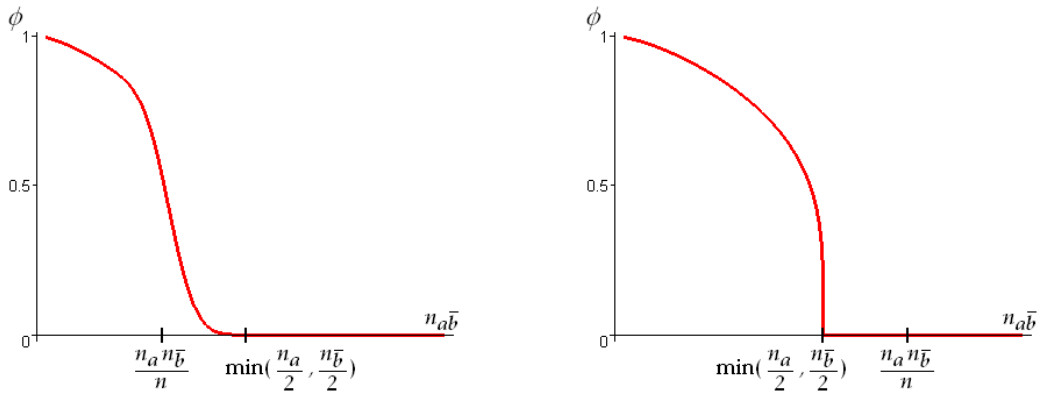
Objet	écart à l'équilibre
Portée	quasi-implication
Nature	statistique
Domaine de variation	[0 ; 1]
Valeur pour les règles logiques	$\sqrt{1 - e^{-\frac{n_a n_{\bar{b}}}{n}}}$
Valeur pour les règles à l'équilibre	0
Valeur pour les règles à l'indépendance	$\leq \frac{1}{\sqrt{2}}$

TAB. 3.2 – Propriétés de l'intensité d'implication entropique

à l'équilibre puisqu'elle ne prend une valeur fixe (en l'occurrence 0) qu'à l'équilibre et non à l'indépendance. Comme nous l'avons vu au chapitre 2, un indice de règle ne peut pas mesurer à la fois un écart à l'équilibre et un écart à l'indépendance.

Considérons une règle $a \rightarrow b$. En faisant varier $n_{a\bar{b}}$ avec n_a, n_b, et n fixes, on peut distinguer deux comportements différents pour l'intensité d'implication entropique :

- Si $n_b \geq \frac{n}{2}$ et $n_a \leq \frac{n}{2}$, alors l'indépendance est atteinte avant les équilibres de la règle et de sa contraposée quand $n_{a\bar{b}}$ augmente. L'intensité d'implication entropique décroît progressivement avant de s'annuler (figure 3.5.(A)).
- Sinon, l'équilibre de la règle et/ou celui de sa contraposée est atteint



(A) $n_b \geq \frac{n}{2}$ et $n_a \leq \frac{n}{2}$ (B) $n_b < \frac{n}{2}$ ou $n_a > \frac{n}{2}$

FIG. 3.5 – Représentation de l'intensité d'implication entropique selon $n_{a\bar{b}}$

Objet	écart à l'équilibre
Portée	règle au sens strict
Nature	statistique
Domaine de variation	[0 ; 1]
Valeur pour les règles logiques	$1 - \frac{1}{2^{n_a}}$
Valeur pour les règles à l'équilibre	0.5
Valeur pour les règles à l'indépendance	< 1

TAB. 3.3 – Propriétés de IPEE

avant l'indépendance quand $n_{a\bar{b}}$ augmente. L'intensité d'implication entropique décroît rapidement et s'annule au premier des deux équilibres, c'est-à-dire lorsque $n_{a\bar{b}} = \min(\frac{n_a}{2}, \frac{n_{\bar{b}}}{2})$ (figure 3.5.(B)).

L'intensité d'implication entropique est de *nature statistique* et varie avec la dilatation des effectifs. Cependant, elle reste discriminante quand les cardinaux étudiés sont grands (voir figure 3.6).

3.5.3 Propriétés d'IPEE

IPEE est un indice de règle au sens de la définition 2.1.3. Ses principales propriétés sont décrites dans le tableau 3.3. L'indice est représenté en fonction du nombre de contre-exemples dans la figure 3.7, et comparé aux principaux indices d'écarts à l'équilibre : la confiance, la moindre-contradiction, et l'in-

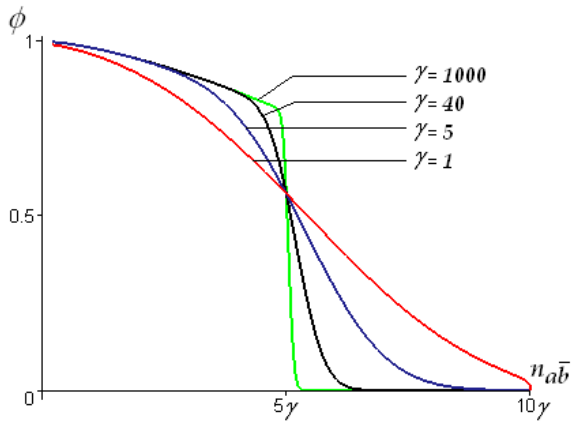


FIG. 3.6 – Représentation de l'intensité d'implication entropique avec la dilatation des effectifs
($n_a = 20 \times \gamma,\ n_b = 75 \times \gamma,\ n = 100 \times \gamma,$
$n_{a\bar{b}} \in [0 \times \gamma\ ;\ 10 \times \gamma],\ \gamma \in \{1;\ 5;\ 40;\ 1000\}$)

dice d'inclusion. Nous pouvons voir que :

- IPEE réagit faiblement aux premiers contre-exemples (décroissance lente). Ce comportement est intuitivement satisfaisant pour un indice statistique puisqu'un faible nombre de contre-exemples ne saurait remettre en cause la règle [85].
- Le rejet des règles s'accélère dans une zone d'incertitude autour de l'équilibre $n_{a\bar{b}} = \frac{n_a}{2}$ (décroissance rapide).

Dans les figures 3.8, les effectifs des données sont multipliés par un coefficient γ à partir d'une configuration initiale. Les indices sont représentés en fonction de γ. Ces figures montrent qu'à proportion exemples/contre-exemples fixée, les indices sont constants sauf IPEE dont les valeurs sont d'autant plus extrêmes (proches de 0 ou 1) que n_a est grand[1]. En effet, de par sa nature statistique, l'indice prend en compte la taille des phénomènes étudiés : plus n_a est grand, plus on peut avoir confiance dans le déséquilibre exemples/contre-exemples observé dans les données, et plus on peut confirmer (figure 3.8.(A)) ou infirmer (figure 3.8.(B)) la bonne qualité de la règle. En particulier, pour IPEE, la qualité d'une règle logique dépend de n_a (voir tableau 3.3). Ainsi IPEE a l'avantage de ne pas attribuer systématiquement la même valeur aux règles logiques. Ceci permet de différencier et hiérarchiser les règles logiques. Parmi les indices d'écart à l'équilibre (voir tableau

[1]Quand la modélisation retenue est gaussienne, ce comportement est visible directement sur $\widetilde{n}_{a\bar{b}}$: $\widetilde{n}_{a\bar{b}} = \sqrt{n_a}(1 - 2 \times confiance)$.

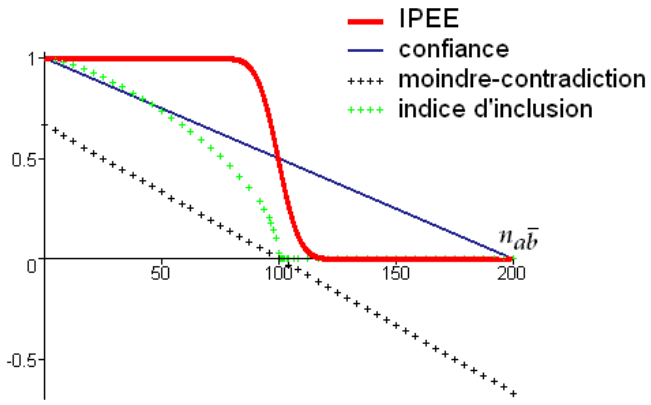
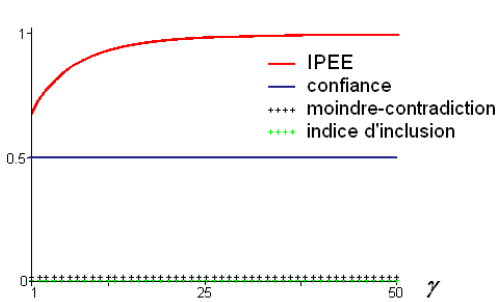
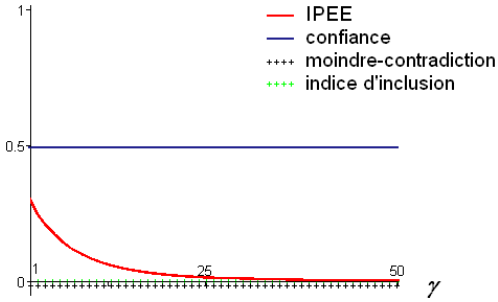


FIG. 3.7 – Représentation des indices d'écart à l'équilibre en fonction de $n_{a\bar{b}}$ ($n_a = 200,\ n_b = 300,\ n = 1000,\ n_{a\bar{b}} \in [0\ ;\ 200]$)



(A) $n_a = 200 \times \gamma,\ n_b = 400 \times \gamma,$ $n_{a\bar{b}} = 97 \times \gamma,\ n = 1000 \times \gamma,$ $\gamma \in [1\ ,\ 50]$



(B) $n_a = 200 \times \gamma,\ n_b = 400 \times \gamma,$ $n_{a\bar{b}} = 103 \times \gamma,\ n = 1000 \times \gamma,$ $\gamma \in [1\ ,\ 50]$

FIG. 3.8 – Représentation des indices d'écarts à l'équilibre en fonction de la dilatation des effectifs

2.2 page 21), seuls la moindre-contradiction et IPEE possèdent cette caractéristique : la moindre-contradiction différencie les règles logiques selon n_b (l'indice favorise les conclusions rares), tandis que IPEE différencie les règles logiques selon n_a (l'indice favorise les prémisses fréquentes).

IPEE porte sur des règles au sens strict et ne possède donc aucune symétrie. On a toutefois la relation suivante :

$$IPEE(a \rightarrow \bar{b}) = 1 - IPEE(a \rightarrow b) - \frac{C_{n_a}^{n_{ab}}}{2^{n_a}}$$

(le dernier terme est négligeable quand n_a est grand)

Comme nous l'avons vu au chapitre 2, les mesures de significativité statistique tendent à être peu discriminantes quand les cardinaux étudiés sont

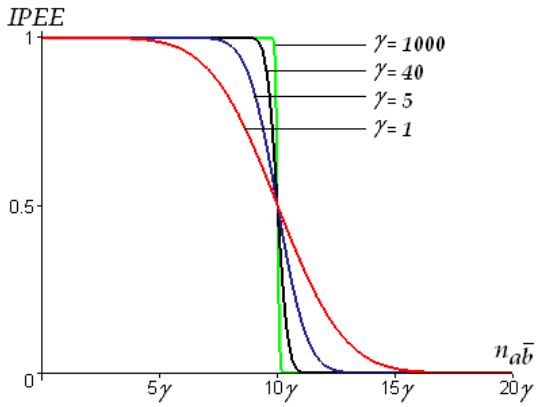


FIG. 3.9 – Représentation de IPEE avec la dilatation des effectifs ($n_a = 20 \times \gamma,\ n_{a\bar{b}} \in [0 \times \gamma\ ;\ 20 \times \gamma],\ \gamma \in \{1;\ 5;\ 40;\ 1000\}$)

grands (de l'ordre de 10^3), car même des écarts faibles peuvent s'avérer statistiquement significatifs au regard d'effectifs importants. Comme l'illustre la figure 3.9, IPEE ne déroge pas à la règle : quand n_a est grand, l'indice tend à évaluer que les règles sont soit très bonnes (valeurs proches de 1), soit très mauvaises (valeurs proches de 0). Dans ce cas, pour affiner le filtrage des meilleures règles, il faut utiliser en supplément de IPEE une mesure descriptive. En revanche, contrairement à l'intensité d'implication ou à l'indice de vraisemblance du lien, IPEE ne dépend pas de n. L'indice est donc autant sensible aux règles spécifiques ("pépites de connaissances") qu'aux règles générales, et a l'avantage d'être adapté à l'étude des petites bases de données comme des grandes.

3.5.4 Propriétés de TI

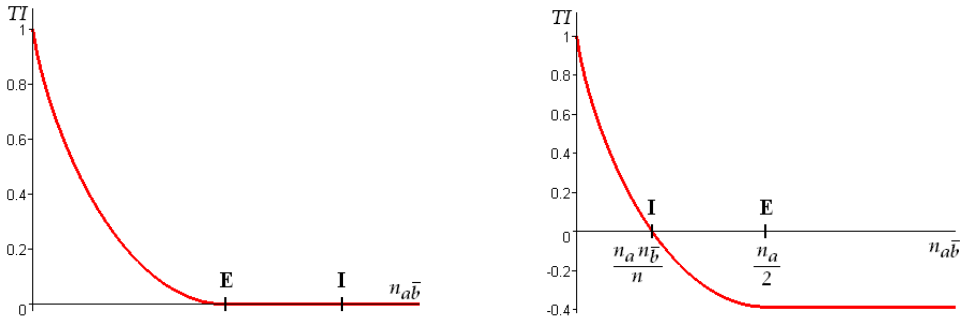
Les principales propriétés de TI sont données dans le tableau 3.4. TI est une fonction décroissante convexe du nombre de contre-exemples. Il fait partie des indices de règle "exigeants" qui diminuent rapidement dès les premiers contre-exemples et permettent ainsi de mieux hiérarchiser les bonnes règles (plus grande dispersion des valeurs). Comme les mesures entropiques, TI est de nature descriptive.

Considérons une règle $a \rightarrow b$. En faisant varier $n_{a\bar{b}}$ avec n_a, n_b, et n fixes, on peut distinguer deux comportements différents pour TI [27] :

- Si $n_b \geq \frac{n}{2}$, alors l'indépendance est atteinte avant l'équilibre quand $n_{a\bar{b}}$ augmente. Le taux informationnel s'annule à l'indépendance puis admet des valeurs négatives (figure 3.10.(A)).
- Si $n_b \leq \frac{n}{2}$, alors l'équilibre est atteint avant l'indépendance quand $n_{a\bar{b}}$

Objet	écart à l'indépendance
Portée	règle au sens strict
Nature	descriptive
Domaine de variation	$]-\infty\ ;\ 1]$
Valeur pour les règles logiques	1
Valeur pour les règles à l'équilibre	$1-\widehat{H}(b)^{-1}\leq 0$
Valeur pour les règles à l'indépendance	0

TAB. 3.4 – Propriétés de TI



(A) $n_b \geq \frac{n}{2}$ (B) $n_b \leq \frac{n}{2}$

FIG. 3.10 – Représentation de TI en fonction de $n_{a\bar{b}}$

augmente. Le taux informationnel s'annule mais n'admet pas de valeurs négatives (figure 3.10.(B)).

A notre connaissance, TI est le seul indice de règle qui puisse rejeter à la fois indépendance et équilibre avec un seuil fixe. C'est une approche tout à fait originale pour l'évaluation de la qualité des règles.

3.5.5 Propriétés de TIC

Les principales propriétés de TIC sont données dans le tableau 3.5. TIC permet de repérer simultanément les situations d'équilibre et d'indépendance pour la règle ou sa contraposée.

Objet	écart à l'indépendance
Portée	quasi-implication
Nature	descriptive
Domaine de variation	[0 ; 1]
Valeur pour les règles logiques	1
Valeur pour les règles à l'équilibre	0
Valeur pour les règles à l'indépendance	0

TAB. 3.5 – Propriétés de *TIC*

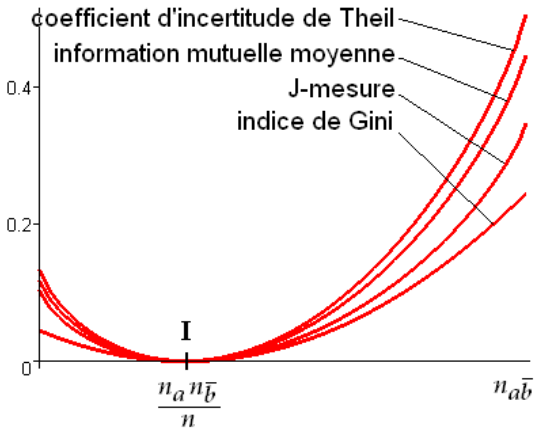
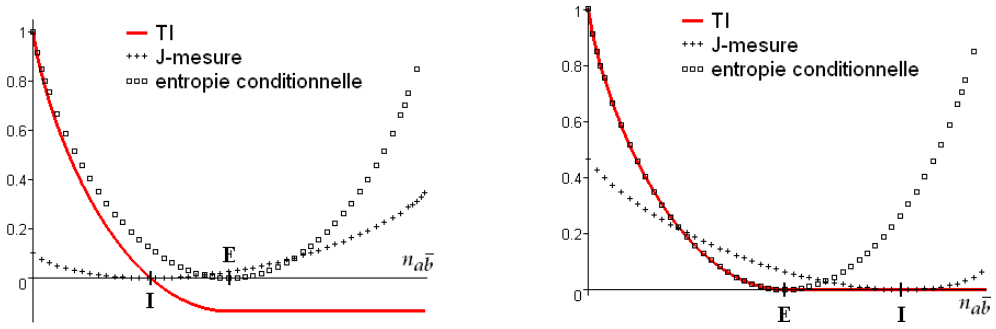


FIG. 3.11 – Les mesures entropiques utilisées pour évaluer des règles



(A) $n_b \geq \frac{n}{2}$ (B) $n_b \leq \frac{n}{2}$

FIG. 3.12 – Représentation de *TI*, la J-mesure, et l'entropie conditionnelle en fonction de $n_{a\bar{b}}$

3.5.5.1 Comparaisons de TIC à d'autres mesures

Comparaisons formelles

Dans cette section, nous comparons le taux informationnel TI aux mesures issues de la théorie de l'information traditionnellement utilisées pour évaluer la qualité des règles : l'entropie conditionnelle de Shannon, l'information mutuelle moyenne, le coefficient d'incertitude de Theil, la J-mesure, et l'indice de Gini. Etant donné que les quatre dernières mesures sont très similaires (voir figure 3.11), nous n'en considérons qu'une seule parmi les quatre dans les comparaisons qui suivent. Nous choisissons la J-mesure, qui est la plus utilisée dans le contexte des règles d'association. En ce qui concerne l'entropie conditionnelle, pour une règle $a \rightarrow b$ ce n'est pas la fonction $H(b/a = 1)$ décrite à la section 3.2.1 qui est représentée dans les comparaisons, mais la fonction complémentaire $1 - H(b/a = 1)$. En effet, contrairement aux autres mesures, $H(b/a = 1)$ attribue ses plus petites valeurs aux meilleures règles (pour générer des règles de qualité, l'algorithme CN2 cherche à minimiser $H(b/a = 1)$ [43]).

Les figures 3.12.(A) et 3.12.(B) comparent TI à l'entropie conditionnelle et à la J-mesure quand le nombre de contre-exemples $n_{a\bar{b}}$ augmente. Les figures illustrent clairement que l'entropie conditionnelle et la J-mesure ne sont pas des indices de règle, puisqu'elles peuvent croître quand les contre-exemples augmentent. De plus, la J-mesure repère l'indépendance (elle s'y annule) mais pas l'équilibre (elle peut même prendre des valeurs élevées à l'équilibre), alors que l'entropie conditionnelle repère l'équilibre (elle s'y annule) mais pas l'indépendance (elle peut même prendre des valeurs élevées à l'indépendance). Dans tous les cas, filtrer les règles sur TI avec un seuil strictement positif suffit pour rejeter à la fois équilibre et indépendance. Comme

	Nombre d'items	Nombre d'individus	Nombre de règles découvertes
T10.I4.D5k	12	5000	97688
T10.I4.D100k	1000	100000	478894
PANNES	92	2883	43930
PROFILS	30	2299	28938

TAB. 3.6 – Caractéristiques des données

l'illustre la figure 3.12.(B), TI est analogue à l'entropie conditionnelle quand $n_b \leq \frac{n}{2}$ (les fonctions sont partiellement identiques). C'est ce qui permet à TI de s'annuler à l'équilibre quand $n_b \leq \frac{n}{2}$.

Comparaisons expérimentales

Nous comparons les distributions de TI aux distributions d'autres mesures sur un ensemble de règles d'association extraites à partir de quatre jeux de données (décrits dans le tableau 3.6). Les deux premiers jeux de données ont été créés à l'aide du générateur[2] de données synthétiques d'IBM décrit dans [5], qui simule des achats dans un supermarché. Les deux autres jeux de données sont une base de données de pannes d'ascenseurs fournie par une société de maintenance, et une base de profils psychologiques utilisée en gestion des ressources humaines, appartenant à la société *PerformanSe SA*[3]. Les règles ont été extraites à l'aide de l'algorithme *Apriori* [5] avec un seuil de support faible pour éviter l'élimination prématurée de règles potentiellement intéressantes (voir chapitre 3 pour plus de détails sur l'algorithme *Apriori*).

Puisque nous souhaitons ici comparer les distributions des mesures, nous choisissons des mesures qui, comme TI, ont 1 pour valeur maximale. Parmi les mesures entropiques, seule l'entropie conditionnelle satisfait à cette condition. Nous ajoutons donc à nos comparaisons deux indices de règle qui vérifient cette condition : la confiance et l'indice de Loevinger (voir définitions dans le tableau 2.1 page 14). Ils mesurent respectivement un écart à l'équilibre et un écart à l'indépendance. Comme le montre la figure 3.13, le taux informationnel TI est l'indice le plus filtrant : pour les quatre jeux de données, quel que soit le seuil choisi entre 0 et 1, TI élimine plus de règles que les autres mesures. Ceci est particulièrement utile pour le post-traitement de grands ensembles de règles.

Expliquons pourquoi TI est un indice très filtrant. Dans les figures 3.14 en coordonnées parallèles, chaque ligne brisée représente une règle. La figure

[2]http ://www.almaden.ibm.com/software/quest/Resources/index.shtml
[3]www.performanse.fr

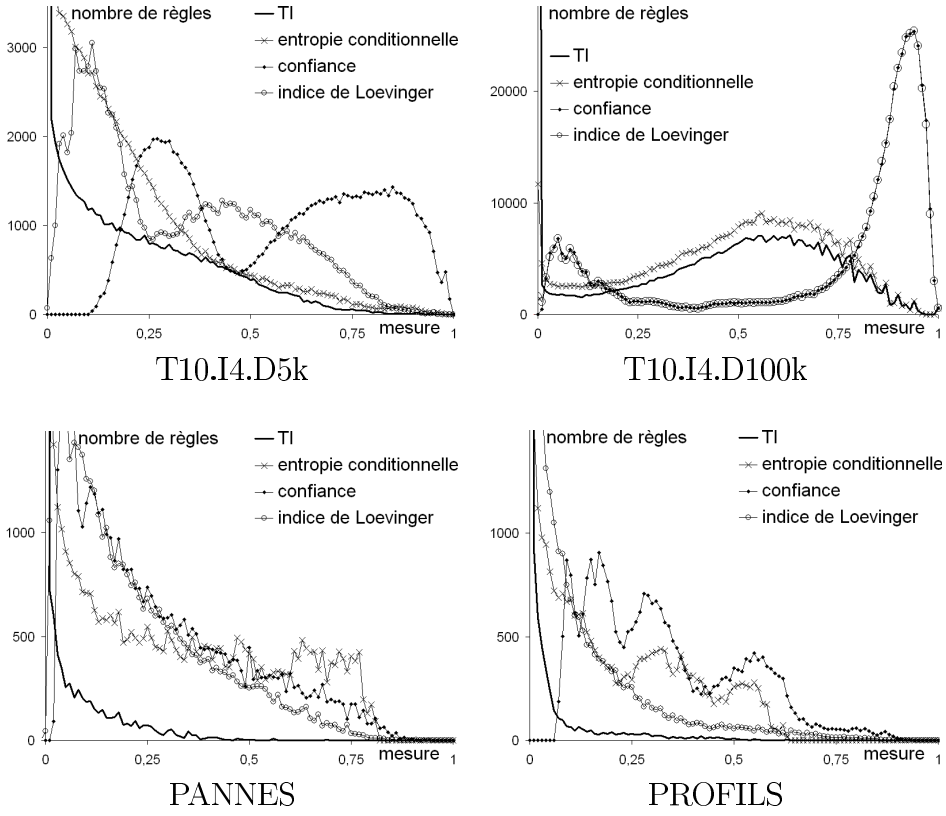
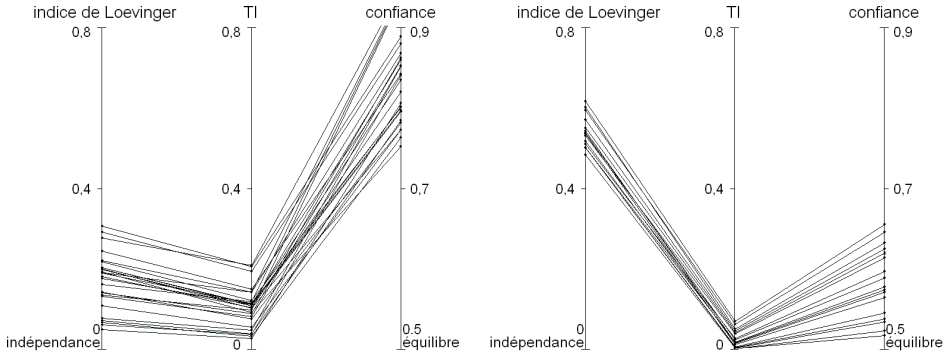


FIG. 3.13 – Distributions des mesures sur les ensembles de règles



(A) échantillon de règles extrait de T10.I4.D5k

(B) échantillon de règles extrait de PANNES

FIG. 3.14 – Deux échantillons de règles représentés en coordonnées parallèles

3.14.(A) montre des règles représentatives de T10.I4.D5k qui sont jugées bonnes par la confiance mais pas par l'indice de Loevinger, alors que la figure 3.14.(B) exhibe des règles de PANNES qui sont jugées bonnes par l'indice de Loevinger mais pas par la confiance. En prenant en compte équilibre et indépendance, TI donne de mauvaises valeurs à toutes ces règles.

3.6 Synthèse des publications

Partant l'analyse statistique implicative [83], et des travaux menés sur l'intensité d'implication, nous avons étudié et proposé un ensemble conséquent d'extensions.

En premier lieu, trois mesures originales et complémentaires ont été conçues :

1. L'*intensité d'implication entropique* (IIE) [31] [30] [29], une extension entropique de l'intensité d'implication adaptée aux données volumineuses et prenant en compte la contraposée ;
2. L'*indice probabiliste d'écart à l'équilibre* (IPEE) [25] [23] [24] [16], la seule mesure statistique (probabiliste) sensible à l'écart à l'équilibre ;
3. Le *taux informationnel modulé par la contraposée* (TIC) [26] [27] [16], une mesure entropique non probabiliste sensible simultanément aux écarts à l'équilibre et à l'indépendance, et prenant en compte la contraposée.

Ces trois nouvelles mesures ont été étudiées et comparées de manière détaillée dans la thèse de Julien Blanchard [16].

En complément, cinq autres extensions de la mesure d'intensité d'implication ont été conçues en dehors du contexte des règles d'association : (i) l'intensité d'implication ordinale, qui propose une extension probabiliste aux *variables numériques* [94][95] ; (ii) une adaptation aux règles sur des *séquences d'événements* [17], appliquée aux pannes d'ascenseurs ; (iii) une mesure probabiliste de quasi-équivalence associée à un algorithme de *réduction des variables redondantes* [89] [48] ; (iv) une mesure de *typicalité* d'un individu (enregistrement) pour une règle [87] ; et enfin (v) une adaptation aux *variables floues* [86].

Parallèlement, nous nous sommes aussi intéressé à l'évaluation de la qualité d'ensembles de règles d'association. Dans ce cadre nous avons proposé une méthode d'*élimination des règles d'association redondantes* [154][153] inspirée des travaux sur les dépendances fonctionnelles.

Enfin, nos travaux ont fait l'objet d'applications à la validation de données psychométriques [56] [57] [58].

Chapitre 4

Graphes de corrélation

Sommaire

Le choix des mesures de qualité, ou d'intérêt, afin d'évaluer les règles d'association est devenu une question importante pour le post-traitement en ECD. Dans la littérature, de nombreux auteurs ont discuté et comparé les propriétés des mesures (cf. chapitre 2) afin d'améliorer le choix des meilleures. Cependant, ces études théoriques atteignent leurs limites, car il s'avère que la qualité d'une règle est contextuelle : elle dépend à la fois de la *structure de données*, et de leur usage c'est-à-dire des *buts* du décideur. Ainsi, certaines mesures peuvent être appropriées dans un certain contexte, mais pas dans d'autres.

Dans ce chapitre, nous présentons une nouvelle approche contextuelle mise en application par un nouvel outil, ARQAT, permettant à un décideur d'évaluer et de comparer expérimentalement le comportement des mesures objectives sur ses jeux de données spécifiques. Cette approche, de type aide à la décision, est basée sur l'analyse visuelle d'un graphe de corrélation entre des mesures. Elle permet in fine de guider l'utilisateur vers le choix des meilleures mesures et ainsi d'identifier les meilleures règles de son corpus.

Après avoir introduit quelques notations et définitions afin de préciser la nature des données traitées, nous présentons un extrait des résultats d'une analyse exploratoire de ces données obtenus par l'utilisation de techniques de classification. Puis nous proposons une nouvelle approche de visualisation par des graphes de corrélation, déclinés en 3 types de graphes. Ces graphes facilitent la visualisation des liaisons entre mesures, permettent d'effectuer des études comparatives, et de détecter des régularités corrélatives. Enfin, nous employons ensuite cette approche afin de comparer et de discuter le comportement de 36 mesures d'intérêt sur deux ensembles de données a priori très opposés : un premier dont les données sont fortement corrélées et un second aux données faiblement corrélées. Alors que nous attendions des différences importantes entre les graphes de corrélation de ces 2 jeux d'essai, nous avons pu observer des stabilités de corrélation entre certaines mesures qui sont révélatrices de régularités indépendantes de la nature des données observées. Ce chapitre s'achève sur une synthèse des publications et des travaux menés sur ce thème.

4.1 Corrélation et notations

Dans ce chapitre, nous nous limitons aux mesures qui sont fonction des effectifs $(n, n_a, n_b, n_{a\bar{b}})$ associés à une règle $a \rightarrow b$, et nous supposerons que pour chaque règle $r{=}a \rightarrow b$ nous disposons de ces effectifs.

Soient $R = \{r_1, r_2, ..., r_p\}$ un ensemble de p règles d'association extraites d'un ensemble de données D. Et soit $M = \{m_1, m_2, ..., m_q\}$ un ensemble de q mesures d'intérêt.

Nous pouvons donc calculer la matrice $MR = ((m_{ij}))$, où $m_{ij} = m_i(r_j)$ correspond à la valeur de la mesure m_i pour la règle pour la règle r_j.

Nous proposons ensuite de calculer la matrice de corrélation entre mesures $MM = ((\rho_{ij}))$, où $\rho_{ij} = \rho(m_i, m_j)$ évalue la corrélation entre les valeurs des mesures m_i et m_j qui correspondent aux colonnes i et j de la matrice MR. Plus précisément, si nous notons $m_i(R) = \{m_{i1}, m_{i2}, ..., m_{ip}\}$ le vecteur

correspondant à la ième colonne de MR, la corrélation sera calculable à partir d'une formule du type $\rho_{ij} = f(m_i(R), m_j(R))$.

Nous avons mené les calculs avec 3 variantes de l'indice de corrélation [195] : La corrélation linéaire (ou corrélation de Pearson) , la corrélation de rang de Spearman, la corrélation de rang de Kendall.

Ainsi la matrice MR permet d'analyser les liens entre les règles et les mesures, et la matrice MM les liens entre les mesures. Cette partie ayant pour objectif de mieux comprendre le comportement des indices, nous avons ciblé notre étude sur la matrice MM.

En pratique, comme la corrélation est symétrique, seule une demi matrice MM contenant $q(q-1)/2$ valeurs de corrélation doit être stockée.

D'autre part, il est possible de transformer la matrice de corrélation MM en une matrice de similarité S (cf [125]).

Afin d'interpréter la matrice de corrélation MM, nous introduisons deux seuils de significativité τ et θ pris dans l'intervalle réel $[0, 1]$ et associés aux deux définitions suivantes :

- **Mesures corrélées (τ-corrélation).** Deux mesures m_i et m_j sont τ-corrélées si : $|\rho(m_i, m_j)| \geq \tau$. Généralement le seuil τ est proche de 1.
- **Mesures non-corrélées (θ-noncorrélation).** Deux mesures de m_i et m_j sont θ-noncorrélées si : $|\rho(m_i, m_j)| \leq \theta$. Généralement le seuil τ est proche de 0.

Pour fixer le seuil de θ-noncorrélation, nous le utilisons le seuil de significativité ([193]) $\theta = 1.960/\sqrt{p}$ où p est le nombre de règles. Les valeurs communes pour θ sont : $\theta = 0.1, 0.05, 0.005$, Nous choisirons $\theta = 0.05$. Nous fixons le seuil de τ-corrélation à $\tau = 0.85$ qui est une valeur commune dans la littérature.

4.2 Résultats expérimentaux obtenus par classification

Afin de faciliter l'analyse de la matrice de corrélation MM par un utilisateur, nous avons choisi de nous orienter vers des techniques de visualisation.

Dans un premier temps, nous avons fait appel aux méthodes d'*analyse exploratoire des données*, qui ont l'avantage d'être visuelles et accessibles (à travers l'outil R notamment).

Partant de la matrice S de similarité entre mesures, nous avons mené une série d'études à l'aide de méthodes de classification classiques. A travers ces

études nous proposons de :

- rechercher des clusters de mesures,
- identifier la meilleure mesure de chaque cluster,
- comparer les clusters sur plusieurs jeux de données.

A cette fin, nous avons utilisé trois méthodes de classification : la classification ascendante hiérarchique (CAH), la classification par k-moyennes (k-means), la classification par médioïdes (PAM). Partant d'une matrice de similarité issue d'un corpus de 35 mesures et de 100000 règles, nous avons pu montrer dans **[125]**, [130], [129], [127] l'intérêt de ces méthodes de classification pour :

- interpréter visuellement les liaisons entre indices ;
- identifier des clusters mettant en évidence des mesures similaires, et ainsi permettre à l'utilisateur de ne pas utiliser des mesures trop semblables ou redondantes ;
- identifier le meilleur représentant de chaque cluster, et ainsi proposer à l'utilisateur un meilleur sous-ensemble d'indices non redondants.

A titre d'illustration la figure 4.1 présente le dendrogramme des similarités entre mesures résultant d'une CAH, et la figure 4.2 les clusters obtenus par la méthode PAM (extrait de [125]).

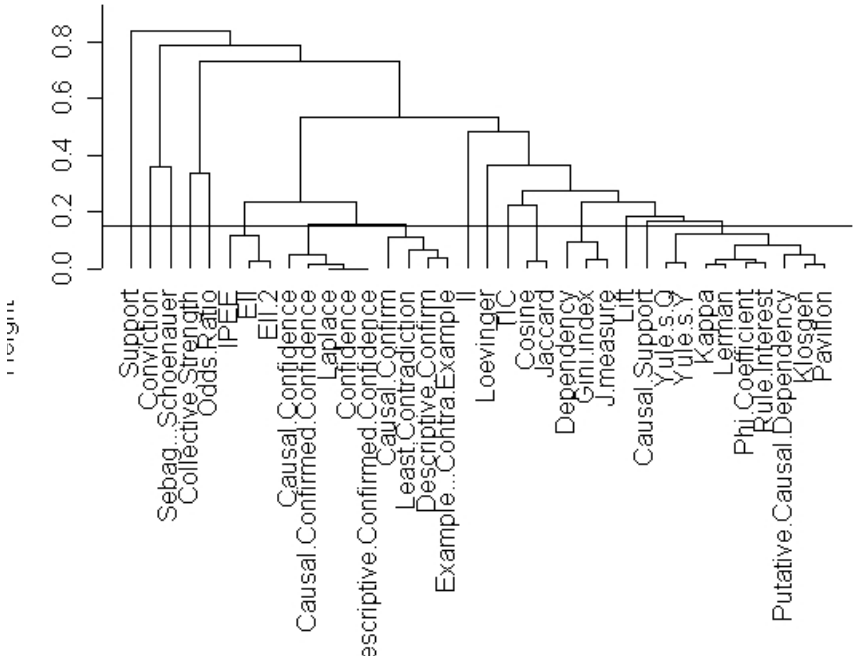


FIG. 4.1 – CAH de 35 mesures de corrélation (Mushrooms).

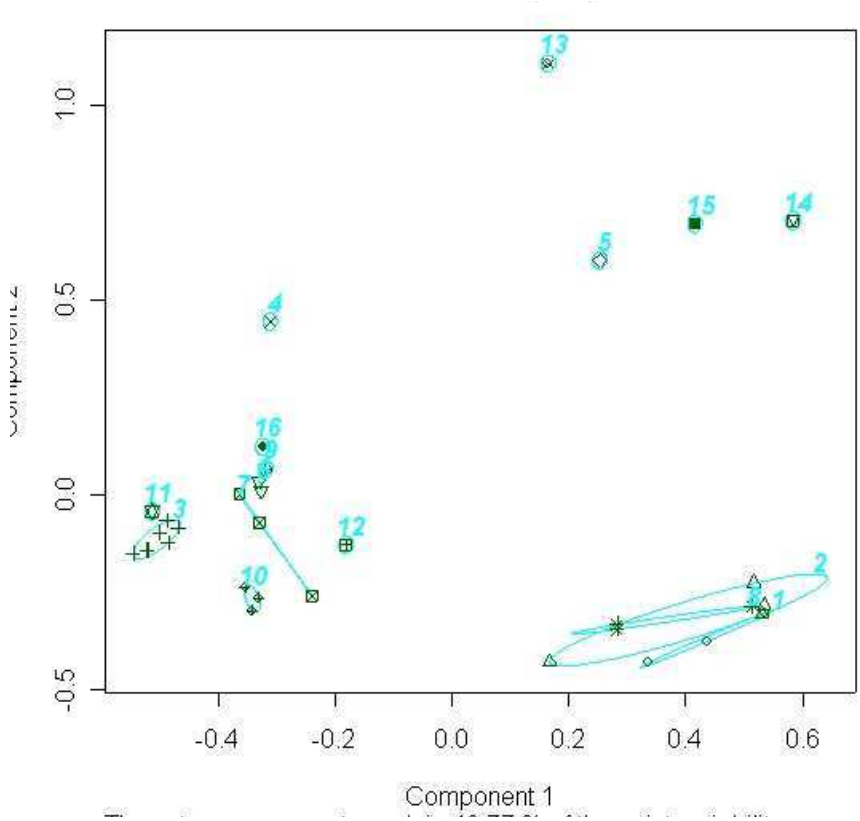


FIG. 4.2 – Projection sur les 2 facteurs principaux de la classification par Médioïdes (Mushrooms).

4.3 Graphes de Corrélation

Malgré l'intérêt des résultats obtenus par les méthodes de classification, nous nous sommes heurtés à deux limites. La première provient de la difficulté rencontrée par des non-spécialistes à l'utiliser convenablement ces techniques. La deuxième est liée à notre besoin d'effectuer des études comparatives entre plusieurs corpus afin de détecter les stabilités et les fluctuations corrélatives, besoin auquel nous n'avons pas trouvé de réponse convenable par l'utilisation des ces méthodes (cf. [127]).

Nous avons donc choisi de nous orienter vers une visualisation à base de graphes qui nous a semblé mieux adaptée à nos objectifs.

4.3.1 Principe

La matrice de corrélation MM peut également être vue comme la relation d'un graphe non-orienté et valué, appelé *graphe de corrélation* (voir Fig. 4.3). Dans ce graphe un sommet est une mesure d'intérêt m_i et une arête est valuée par la valeur de corrélationρ_{ij} entre deux sommets/mesures m_i et m_j.

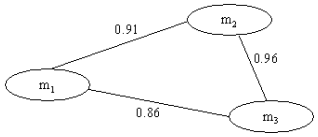


FIG. 4.3 – Une illustration de graphe de corrélation.

Par exemple, la figure Fig. 4.3 montre un graphe de corrélation obtenu sur cinq règles d'association $R(D) = \{r_1, r_2, r_3, r_4, r_5\}$ et trois mesures $M = \{m_1, m_2, m_3\}$ dont les valeurs et les corrélations sont indiquées dans le Tableau Tab. 4.1.

$R(D)$	m_1	m_2	m_3
r_1	0.84	0.89	0.91
r_2	0.86	0.90	0.93
r_3	0.88	0.94	0.97
r_4	0.94	0.95	0.99
r_5	0.83	0.87	0.84

CC	m_1	m_2	m_3
m_1		0.91	0.86
m_2			0.96
m_3			

TAB. 4.1 – Valeurs de corrélation pour trois mesures et cinq règles d'association.

4.3.2 Graphe des mesures corrélées *(CG+)* versus non-corrélées *(CG0)*

Malheureusement, le graphe de corrélation issu de la matrice de corrélation MM est complet, et n'est donc pas directement exploitable par l'utilisateur. Afin de résoudre ce problème, nous proposons d'utiliser la τ-corrélation et la θ-noncorrélation (cf section 4.1) pour définir deux transformations qui permettront d'extraire deux sous-graphes plus limités et plus lisibles.

En premier lieu, nous pouvons extraire *le sous-graphe corrélé partiel (CG+)* : la partie du graphe où nous ne retenons que des arêtes liées à une forte corrélation (τ-corrélation). En second lieu, nous pouvons construire le *sous-graphe non-corrélé partiel (CG0)* où nous ne retenons que les arêtes liées aux valeurs de corrélation proches de 0 (θ-noncorrélation).

S'ils sont bien dessinés, ces deux sous-graphes partiels $CG+$ et $CG0$ pourront être utilisés comme support de visualisation immédiat et intelligible afin d'observer les liaisons corrélatives entre mesures.

La figure 4.4 montre les deux graphes $CG+$ et $CG0$ extraits de la base de données mushroom (extrait de [124]).

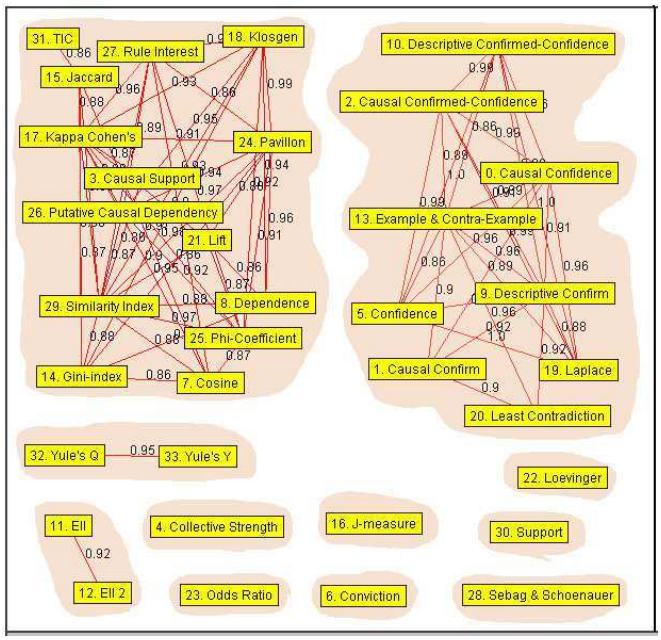


Figure 4.4: Graph CG+ des τ-corrélations.

On peut également y observer des clusters d'indices corrélés correspondant aux parties connexes de ces graphes.

4.3.3 graphe de stabilité

Afin de comparer les corrélations de mesures entre plusieurs jeux de données, nous introduisons une extension des graphes de corrélation appelée *graphe de stabilité* .

- **Graphes de stabilité.** Le graphe τ-stable $\overline{CG+}$ (resp. θ-stable $\overline{CG0}$) d'un ensemble de k jeux de règles $R = \{R_1, ..., R_k\}$ est défini comme le graphe intersection des k graphes $CG+$ (resp. $CG0$) calculés sur R. Chaque arête retenue est alors valuée par $\overline{\rho}_{ij}$ la moyenne des k corrélations ρ_{ij} observées entre les mesures m_i et m_j sur les k jeux de règles. Ainsi le graphe τ-stable $\overline{CG+}$ (resp. θ-stable $\overline{CG0}$) permettra de visualiser les fortes corrélations (resp. non-corrélations) stables, comme

étant communes aux k jeux de données étudiés. Leurs complémentaires donneront les corrélations instables.

– **Corrélations τ-stable (resp. θ-stable).** On appellera corrélations τ-stable (resp. θ-stable) les τ-corrélations (resp. θ-corrélations) du graphe $\overline{CG+}$ (resp. θ-stable du graphe $\overline{CG0}$).

La figure 4.4 montre les graphes de stabilité obtenus sur 4 jeux de règles valués par 36 mesures (extrait de [126]).

4.4 Résultats expérimentaux avec les graphes de corrélation

Nous avons réalisé une *étude comparative* sur deux jeux de données a priori très différents : la base de données Mushroom (D_1) dont les variables sont fortement corrélées, et à l'opposé une base de données aléatoire (D_2) issue d'un générateur automatique et dont les variables sont faiblement corrélées. Puis, parmi les règles d'associations R_1 (resp. R_2) extraites de D_1 (resp. D_2), nous distinguons le sous ensemble $R_1^{'}$ (resp. $R_2^{'}$) des règles bien évaluées par au moins une mesure. Le tableau 4.2 récapitule les caractéristiques des 2 jeux de données (D_1, D_2) et des 4 jeux de règles (R_1, R_2, $R_1^{'}$, $R_2^{'}$) sur lesquels nous menons l'étude comparative de 40 mesures d'intérêt en utilisant la corrélation linéaire.

	Nombre de variables	Taille moyenne des itemsets	Nombre de transactions	$R(D)$	Nombre de règles	θ	τ
D_1	118	22	8416	R_1	123228	0.005	0.85
				$R_1^{'}$	10431	0.020	0.85
D_2	81	5	9650	R_2	102808	0.003	0.85
				$R_2^{'}$	7452	0.012	0.85

Table 4.2: Description des données.

4.4.1 Comparaison visuelle des Graphes CG0 et CG+

La figure Fig. 4.5 présente les quatre graphes $CG+$ obtenus, et le tableau Tab. **??** récapitule le nombre des corrélations. Ces 4 graphes synthétisent les corrélations significatives, c'est-à-dire les mesures dont le point de vue sur les données est proche, En observant les graphes, nous pouvons noter 4 résultats :

- il existe de nombreuses corrélations entre les mesures,
- Plusieurs clusters (parties connexes grisées) de mesures similaires apparaissent, et mettent en évidence les groupes de mesures redondantes.
- Les graphes CG+ obtenus sur la totalité des règles ($CG + (R_1)$ et $CG + (R_2)$) et sur les sous-ensembles des meilleures règles ($CG + (R'_1)$ et $CG + (R'_2)$) sont très semblables. Ceci nous indique que sur les 2 jeux de données les corrélations et les clusters formés demeurent stables lorsque l'on sélectionne les meilleures règles.
- Comme on l'attendait, on observe un écart important entre les deux jeux de données, ce qui indique une sensibilité des mesures à la nature des données.
- En revanche, nous pouvons observer un nombre important de corrélations entre mesures sur le jeu de données R_2 - même s'il est 2 fois plus faible que sur R_1 - alors que nous en attendions peu du fait de la faible corrélation des données dans D_2.

La figure Fig. 4.6 permet de visualiser les mesures significativement non-corrélées, dont le point de vue sur les données diffère. On y observe les résultas suivants :

- Il y a un très faible nombre de non-corrélations significatives, ce qui indique que très peu de mesures sont en désaccord fort sur l'évaluation des règles.
- Toutefois, le nombre de mesures non-corrélées augmente lorsque l'on passe de la totalité des règles aux meilleures.
- En revanche, contrairement à ce que l'on pouvait attendre, il y a moins de non-corrélations sur le jeu de données synthétique R_2.
- Enfin, aucun comportement stable n'apparaît entre les mesures sur les 4 graphes $CG0$, et donc le graphe $\overline{CG0}$ est vide.

4.4.1.1 Le graphe $\overline{CG+}$ des corrélations stables

Le résultat le plus surprenant apparaît dans le graphe $\overline{CG+}$. En effet, nous y découvrons 5 clusters de mesures τ-stables, c'est-à-dire dont les corrélations demeurent inchangées entre les jeux de données. Ceci dénote d'une invariance des corrélations avec la nature des données.

En analysant plus précisément ces 5 clusters τ-stable, nous notons quelques éléments intéressants.

- (C1), le plus grand cluster, (Confidence, Causal Confidence, Causal Confirmed-Confidence, Descriptive Confirmed-Confidence, Laplace) rassemble des mesures dérivées de la mesure de confiance (Confidence). De plus, ce lien est fort, puisque le graphe est complet et les valeurs de corrélation supérieures à 0.97. Ceci indique un très fort accord entre ces 5

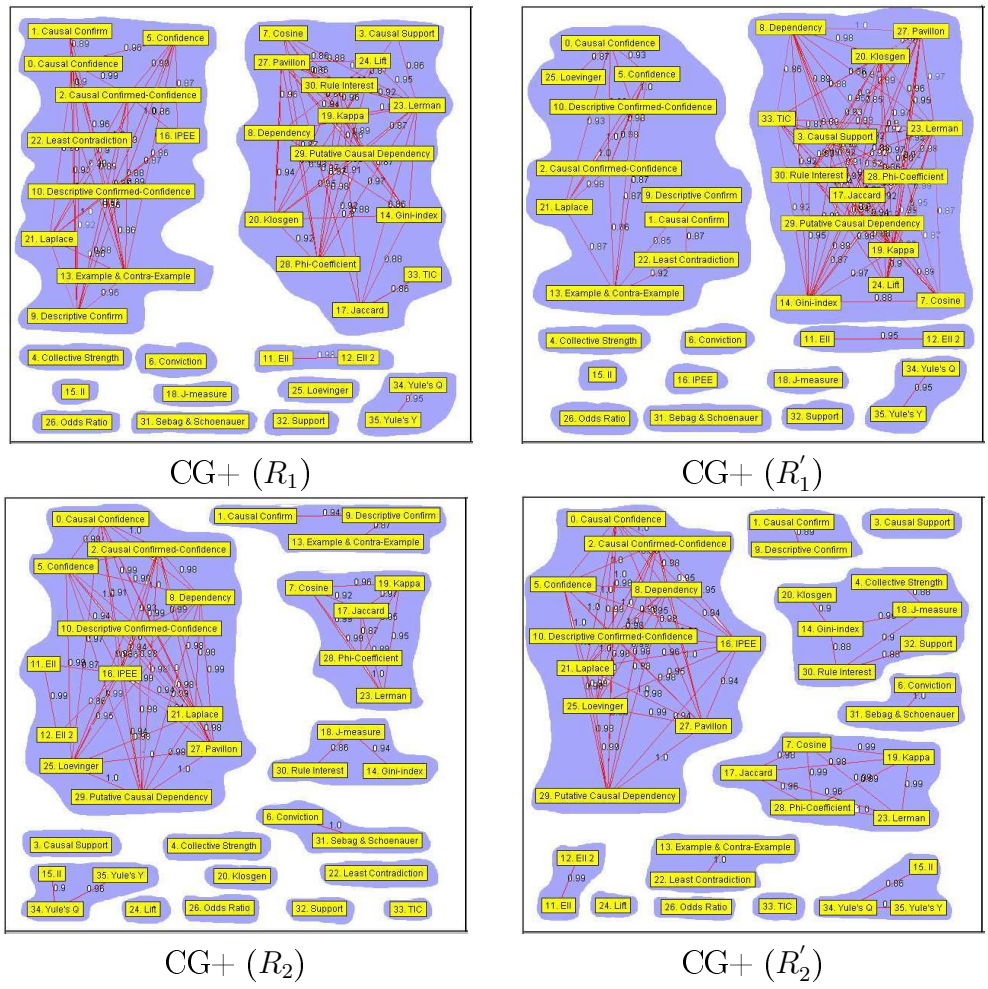
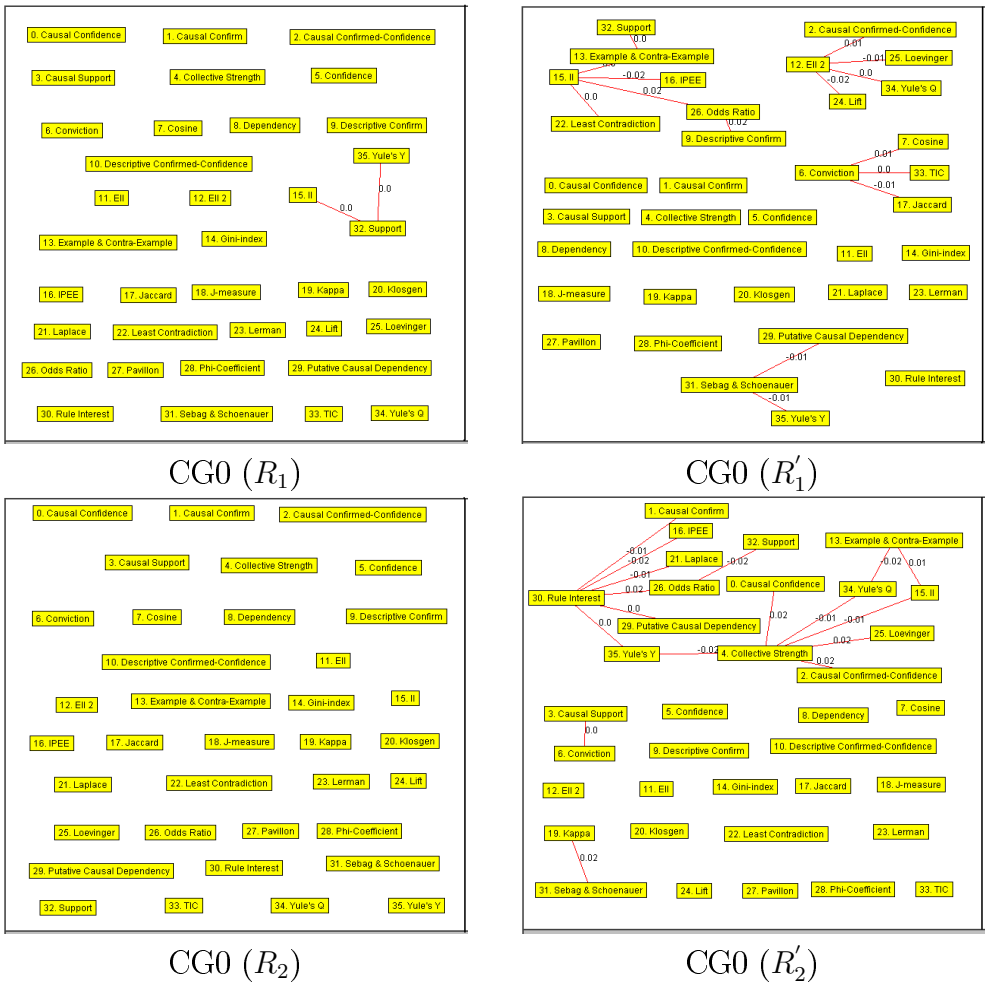


Figure 4.5: Les 4 Graphes CG+ (les clusters sont grisés).

mesures.

- (C2), ce cluster moins fortement corrélé que le premier, est constitué des mesures Phi-Coefficient, Lerman, Kappa, Cosine et Jaccard. Ce cluster rassemble des mesures partageant les 4 propriétés de : symmetric under variable permutation, antisymmetric under row/column permutation, et null invariance [209, 208]. Les deux mesures Jaccard et Cosine ne partagent que la cinquième propriété (null invariance) proposée par [209, 208].
- (C3), rassemble 3 mesures concernées par la première propriété (symmetry/asymmetry under variable permutation) proposée par [209, 208]. L'existence de ce cluster est nécessaire pour distinguer la règle $a \Rightarrow b$ de $b \Rightarrow a$.



CG0 (R_1) CG0 (R_1')

CG0 (R_2) CG0 (R_2')

Figure 4.6: Graphes CG0.

– (C4), est un cluster constitué par deux versions de l'intensité de l'implication entropique EII et EII 2, ce qui n'est pas surprenant.
– (C5), la stabilité de la corrélation Yule'Q et Yule'Y, est elle aussi sans surprise puisque les deux mesures présentent une dépendance fonctionnelle. Ce cluster est lié à la deuxième propriété (row/column scaling invariance) proposée par [209, 208].

Les résultats du graphe τ-stable, donnent une piste intéressante pour construire une base réduite de mesures dont le point de vue est le plus discriminant sur les données. Il suffit pour cela de proposer à l'utilisateur de choisir 5 mesures, une parmi chacun des clusters. Dans [129], nous proposons une solution d'extraction automatique du meilleur représentant de chaque cluster à partir des médioïdes.

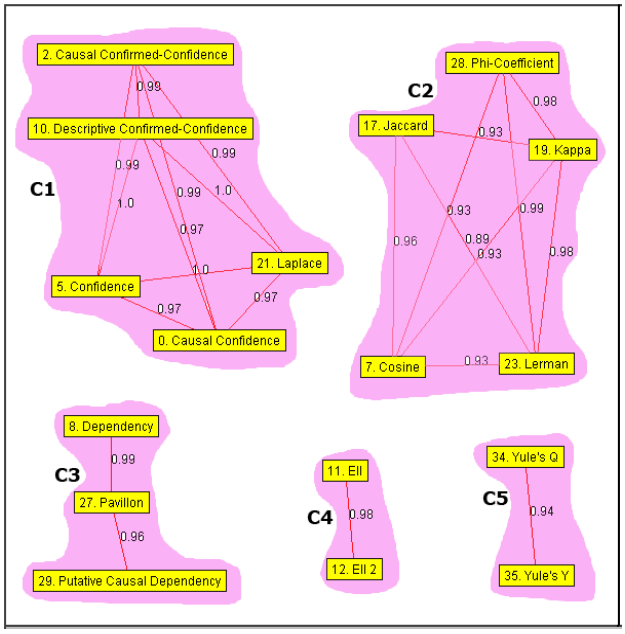


Figure 4.7: Graphe $\overline{CG+}$.

4.5 Outil ARQAT

Afin d'expérimenter et de valider notre approche, nous avons implémenté la plateforme ARQAT (Association Rule Quality Analysis Tool).

Les différentes fonctionnalités d'ARQAT sont décrites de manière détaillée dans [124][121][122][123][120]. La plateforme permet d'étudier le comportement spécifique des mesures d'intérêt sur le jeu de données de l'utilisateur et selon une perspective d'analyse exploratoire. Plus précisément, ARQAT est une boite à outil conçue pour aider graphiquement l'utilisateur analyste à repérer dans ses données les meilleures mesures et au final les meilleures règles.

ARQAT inclut 40 mesures objectives issues d'un recensement bibliographique (cf Chapitre 2).

ARQAT (Fig. 4.8) implémente 14 vues graphiques complémentaires qui sont structurées en 5 groupes selon la tâche offerte.

Les données d'entrée sont constituées d'un ensemble R de règles d'association extrait d'un jeu de données initial D, où la description de chaque règle $a \Rightarrow b$ est complétée par ses contingences $(n, n_a, n_b, n_{a\bar{b}})$ dans D.

Dans une étape préliminaire, l'ensemble de règles R est traité afin de calculer les valeurs des mesures pour chaque règle, puis les corrélations entre chaque paire de mesure. Les résultats sont stockés dans deux matrices : la

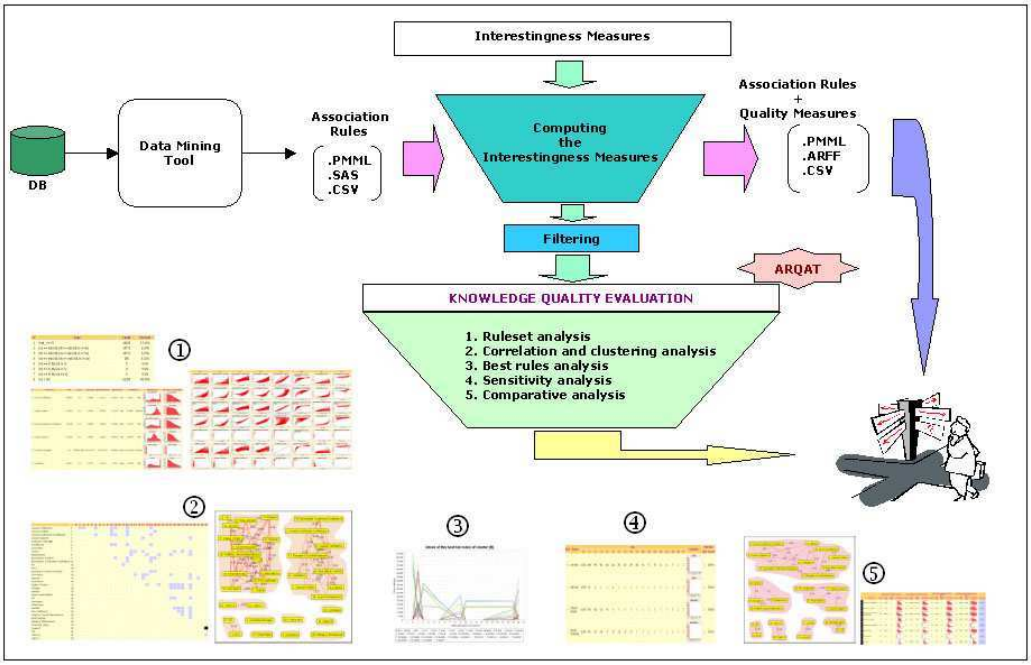


Figure 4.8: Structure d'ARQAT.

matrice MR qui croise règles et mesures, et MM la matrice des corrélations entre mesures.

Lors de cette étape, l'ensemble de règles R peut aussi être échantillonné afin de cibler l'étude sur un sous-ensemble de règles.

La seconde étape est ensuite interactive, l'utilisateur mène l'exploration graphique des résultats. Il s'appuie pour cela sur la structuration en 5 groupes de vues orientées tâche :

- Le premier groupe (1 dans Fig. 4.8) est dédié à la visualisation de statistiques élémentaires afin de mieux appréhender la structure de la matrice MR.
- Le deuxième groupe (2) est orienté vers la visualisation de la matrice de corrélation MM. Il permet la classification des mesures en clusters et l'identification des meilleures mesures.
- Le troisième groupe (3) cible l'extraction des meilleures règles.
- Le quatrième groupe (4) permet une étude de la sélectivité des mesures.
- Enfin, un dernier groupe offre la possibilité de mener une étude comparative des résultats obtenus sur plusieurs ensembles de règles à l'aide des graphes de stabilité.

4.6 Synthèse des résultats obtenus

Nous avons formalisé et proposé une approche visuelle et contextuelle pour faciliter l'analyse des liaisons entre mesures d'intérêt sur des corpus de règles.

Dans un premier temps, nous avons expérimenté trois méthodes de classification des mesures : CAH, k-moyennes et PAM. Nous avons montré dans [125][130] l'intérêt de ces techniques pour la visualisation des liaisons corrélatives et la découverte de clusters. Puis, dans [130][129], nous avons proposé d'extraire les meilleurs représentants de chaque cluster à l'aide de médioïdes. Enfin, dans [127] nous avons mené une étude comparative sur 4 jeux de règles.

Dans un second temps, nous avons proposé une nouvelle approche de visualisation des liaisons entre mesures d'intérêt par des graphes de corrélation. Dans [127], nous avons montré l'intérêt de ce mode original de visualisation par des graphes. Dans [126][128][131][120], nous avons utilisé ces graphes pour mener des études comparatives sur plusieurs corpus de règles de grande taille, et avons proposé de visualiser les liaisons stables par des graphes de stabilité.

Enfin, nous avons implémenté la plateforme ARQAT, Association Rule Quality Tool ([124][121][122][123]), qui intègre un ensemble d'outils d'évaluation des mesures d'intérêt, dont les graphes de corrélation.

Chapitre 5

Fouille de règles

Sommaire

Les algorithmes de fouille de données produisent des règles d'association en si grande quantité que l'utilisateur ne peut généralement pas les exploiter directement. En analyse du panier de la ménagère par exemple, il n'est pas rare d'obtenir des millions de règles portant sur plusieurs milliers d'items. La dernière étape de l'ECD, qui est celle du post-traitement des résultats, se révèle donc particulièrement cruciale dans le contexte des règles d'association : de son efficacité à aider l'utilisateur à explorer cette masse d'information dépend la réussite de tout le processus ECD. Le processus d'ECD glisse ainsi d'une problématique de fouille de données (extraire les régles d'association potentielles cachées dans les données) à celle d'une véritable *fouille de règles*

(extraire les règles d'association utiles cachées parmi les règles d'association potentielles).

Les trois principales approches pour le post-traitement des règles d'association sont les suivantes :

1. évaluer, ordonner, et filtrer les règles avec des indices autres que le support et la confiance ;
2. organiser une exploration interactive des règles pour l'utilisateur ;
3. représenter les règles sous forme graphique.

Les chapitres précédents ont été principalement consacrés au choix des meilleures mesures de qualité (approches 1 et 2). Dans ce chapitre, nous nous intéressons aux choix des meilleures règles (approches 2 et 3).

Dans les travaux dédiés au post-traitement des règles d'association, l'usage de techniques de visualisation est souvent préconisé. Plus généralement, la visualisation est un moyen efficace d'introduire de la subjectivité dans chaque étape du processus ECD [70]. Les représentations visuelles peuvent être exploitées :

- soit en tant que méthode de fouille de données à part entière, ce qui est souvent appelé *visual data mining* [138] ;
- soit en collaboration avec des algorithmes de fouille de données pour faciliter et accélérer l'analyse des données étudiées, des résultats intermédiaires, ou des connaissances produites [1] [199] [112].

Dans ce chapitre, nous proposons une méthodologie pour la visualisation interactive des règles d'association, conçue pour faciliter la tâche de l'utilisateur confronté à de grands ensembles de règles. Elle est fondée sur :

- des principes de visualisation d'information pour la construction de représentations visuelles efficaces [15] ;
- des principes cognitifs de traitement de l'information dans le contexte des modèles de décision [172].

Après un état de l'art sur le post-traitement des règles d'association, nous présentons les principes de notre approche anthropocentrée de fouille de règles d'association fondée sur une stratégie cognitive de *ciblage de règles*. Nous y justifierons les caractéristiques principales de cette approche originale : ses fondements cognitifs, sa représentation graphique et dynamique, le rôle de l'interactivité, et son incidence sur la localité des algorithmes induits. Puis, nous proposons de décliner cette approche en deux visualisations complémentaires couplées à des mesures de qualité :

- Par des *graphes,* en considérant l'ensemble des règles comme un immense réseau dont on ne représente qu'une partie à l'aide d'un graphe. La

principale difficulté posée réside dans la préservation de la carte mentale de l'utilisateur, risquant d'être perturbée par le caractère fortement dynamique de la représentation, puisque celle-ci doit évoluer partiellement à chaque interaction.

- Par des *métaphores 3D* et l'usage de la *Réalité Virtuelle*, afin d'améliorer l'immersion de l'utilisateur dans la représentation. Les principales difficultés se situent dans la synthèse d'une métaphore intelligible et la définition de voisinages de règles.

Ce chapitre s'achève sur une synthèse des publications et des travaux menés sur ce thème.

5.1 Post-traitement des règles d'association : état de l'art

Les approches proposées dans la littératures pour le post-traitement des règles d'association peuvent être classées selon deux tendances : les approches privilégiant l'interactivité avec l'utilisateur, et les approches intégrant une visualisation graphique des règles.

Approches interactives. Les *approches interactives* proposent à l'utilisateur des fonctionnalités de recherche dans un ensemble de règles. Elles constituent un cas particulier de recherche d'information dans des systèmes d'information constitué de règles. On y distingue deux familles de travaux :

- *Les langagues de requête.* Dans le cadre des bases de données inductives [132], plusieurs *langages de requête* ont été développés pour créer et manipuler des règles d'association, comme MSQL [133], DMQL [111], MINE RULE [171], et XMINE [35] .
- *Les explorateurs de règles.* A l'instar des explorateurs de fichiers, les *explorateurs de règles* sont des interfaces interactives qui présentent l'information sous forme textuelle ([139]). Ainsi, dans le logiciel TASA [140], dans [79], et dans [211], l'utilisateur peut définir des contraintes syntaxiques sur les règles recherchées. Dans [167], les auteurs proposent de sélectionner et de spécialiser des règles générales contenues dans un résumé de règles ([162]). Dans [160] et [164], les explorateurs de règles exploitent des indices de règle subjectifs, et peuvent ainsi tirer profit des connaissances de l'utilisateur sur les données.

Approches graphiques. Les *approches de visualisation graphiques* des règles offrent une grande variété de représentations.

- *Par matrices.* Des représentations par *matrice* itemset-à-itemset sont proposées dans [115], [3], DBMiner [109], MineSet[41], Enterprise Miner[1], et DB2 Intelligent Miner Visualization[2]. en donnent différentes implémentations. Cette technique de visualisation a été améliorée en matrices item-à-règle dans [214].
- *Par graphes.* Les règles y sont visualisés à l'aide d'un graphe[3] orienté en 2D (voir [139], [188], et les logiciels DBMiner[2] [109], CHIC[4] [47], et DB2 Intelligent Miner Visualization[5]), et même en 3D, dans [113].
- On trouve aussi des représentations par *mosaïques* ([115]) et par *coordonnées parallèles* ([143], [110]).

Ces approches de post-traitement de règles rencontrent deux limites principales :

- Aucune de ces approches ne combine convenablement interactivité et visualisation. La visualisation est souvent textuelle dans les approches interactives, et les approches de visualisation graphique sont peu interactives.
- Les approches de visualisation graphique se heurtent à la très grande quantité de règles à représenter et deviennent rapidement inexploitables.

5.2 Notre approche de fouille anthropocentrée

5.2.1 Contraintes cognitives

Les travaux en sciences cognitives montrent que la prise de décision est régie par des principes de *rationalité limitée* [201] et de recherche de *structure de dominance* [172], qui contraignent un décideur à ne pouvoir traiter simultanément que de petites quantités d'information. De manière complémentaire, considérant la masse d'information que doit traiter l'utilisateur lors d'un processus d'ECD, [9] propose d'utiliser une méthodologie d'*attribute focusing* qui consiste à cibler de petits groupes d'attributs afin de permettre à l'utilisateur d'interpréter plus facilement les données. L'intérêt de cibler un petit nombre d'attributs pour le traitement cognitif de grandes quantités

[1] www.sas.com/technologies/analytics/datamining/miner/

[2] www.ibm.com/software/data/iminer/visualization/index.html

[3] Pour des règles de plus de deux items, il s'agit en fait d'un hypergraphe : les arcs peuvent contenir plusieurs branches pour relier plusieurs items en prémisse à plusieurs items en conclusion.

[4] www.ardm.asso.fr/CHIC.html

d'information a également été souligné par les travaux de [10] sur l'heuristique de la base mobile.

Afin de mieux intégrer les contraintes cognitives de l'utilisateur dans un processus de fouille, et ainsi d'aller vers un nouveau modèle de fouille antropocentrée de règles d'association, nous proposons d'établir 2 principes directeurs :

- (P1) La quantité de règles représentées doit être réduite (attribute focusing) et intelligible afin de pouvoir rapidement y des règles intéressantes (structure de dominance)
- (P2) En contrepartie l'utilisateur doit pouvoir modifier facilement les règles représentées jusqu'à trouver celles qui l'intéressent.

Ces principes peuvent également être associés aux principes de parcimonie (P2), et de décidabilité (P3) proposés par [10].

5.2.2 Stratégie interactive de *ciblage de règles (Rule Focusing)*

Notre approche de fouille de règles combine les caractères de visualisation et d'interactivité développés précédemment (section 5.1), et permet à l'utilisateur de mener une stratégie de fouille de règles, appelée *ciblage de règles* (ou *rules focusing*), à travers une interface graphique interactive. Notre approche est étagée en 5 niveaux d'abstractions successifs :

- Un *utilisateur* dirigeant la fouille (niveau décisionnel)
- Une *représentation graphique* limitée à un sous ensemble de règles (niveau graphique interactif)
- Des *opérateurs* de manipulation de la représentation (niveau tâche)
- Un *algorithme de fouille* permettant d'extraire des sous ensembles de règles dans les données (niveau fonctionnel)
- Une *base de données* contenant les données à traiter (niveau données)

L'utilisateur mène sa stratégie de fouille dans l'ensemble R de toutes les règles en interagissant exclusivement avec le sous-ensemble S qui lui proposé dans la *représentation graphique*. Le processus interactif se déroule alors de la manière suivantes, à l'étape k : (1) l'*utilisateur* rétroagit sur le sous ensemble de règles S_k proposé dans la *représentation graphique* des via des *opérateurs* d'interaction intégrés à la représentation. (2) Les *opérateurs* d'interaction déclenchent l'appel à l'*algorithme* de fouille. (3) L'*algorithme* lié à l'opérateur extrait dans la *base de données* un nouveau sous-ensemble de règles S_{k+1}. (4) La représentation graphique est alors mise à jour pour intégrer S_{k+1}. Ainsi, tel qu'illustré dans la figure 5.1, l'utilisateur dirige chaque étape de la

construction d'une suite de sous-ensembles de règles $S_0 \rightarrow S_1 \rightarrow ... \rightarrow S_k$ jusqu'à obtenir ce qu'il recherche (principe P2). Chaque transition ($\rightarrow$), liée à un opérateur régi par le principe P1, est déclenchée par l'utilisateur en fonction ses préférences et intègre donc une dimension subjective. La stratégie de rule focusing permet donc à l'utilisateur de procéder par tâtonnements successifs.

De plus, l'algorithme déclenché par chaque opérateur a la particularité d'être *local*. D'un point de vue recherche opérationnelle, l'utilisateur joue donc le rôle d'une heuristique dirigeant une suite d'optimisations locales jusqu'à obtenir une solution acceptable. Cette caractéristique de localité a une conséquence fondamentale sur l'*efficacité algorithmique* de la fouille lors d'un passage à l'échelle. Effet notre approche restera praticable là où les approches globales s'effondrent, en particulier lorsque le nombre d'attributs est grand.

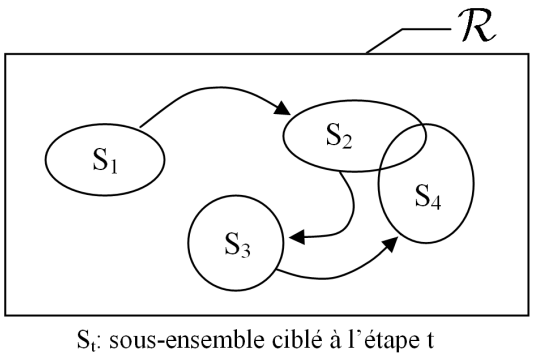


FIG. 5.1 – Des explorations locales successives dans l'ensemble $\mathcal{R}$ des règles

5.3 Visualisation de réseaux de règles

5.3.1 Représentation par des graphes

Ce premier modèle considère les règles extraites au cours du processus, comme un réseau organisé de règles présenté sous la forme d'un graphe orienté. De plus, afin de répondre aux exigences d'interactivité avec l'utilisateur, ce graphe a la particularité d'être dynamique, et d'être manipulable par un ensemble opérateurs permettant à l'utilisateur d'orienter son exploration selon ses centres d'intérêt.

Plus particulièrement, un ensemble de règles d'association est représenté par un graphe orienté sans circuit G = (V, E), dont les sommets de V sont des itemsets et les arcs de E des règles d'association significatives, au sens d'un critère C (défini à partir de mesures d'intérêt, comme le support et la confiance, voir les mesures décrites au chapitre 2). Plus précisément, un arc

de l'itemset X vers l'itemset Y porte la règle d'association $X \rightarrow Y \setminus X$ et présuppose que $X \subset Y$.

La figure 5.2 donne un exemple de graphe de règles portant sur un ensemble d'items {a,b,c,d,e}. Ce graphe décrit l'ensemble des règles d'association suivantes : a→b, a→c, a→e, ab→c, ac→b, ab→e, ae→c. En particulier, la règle ab→c est portée par l'arc partant du sommet a∧b vers le sommet a∧b∧c, et son critère de significativité C(ab→c) est vérifié. Sur cet exemple, nous n'avons représenté que les *règles simples*, celles qui concluent vers un item et sont caractérisées par $|X| = |Y| + 1$.

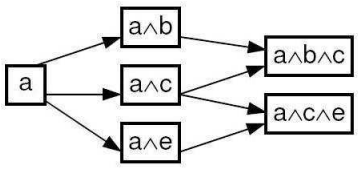


FIG. 5.2 – Graphe de réseau de règles

5.3.2 Dynamique

5.3.2.1 Opérateurs d'interaction

Les opérateurs permettent la manipulation du graphe, afin de réaliser le processus de fouille.

Au départ, le graphe n'est constitué que d'un ensemble d'itemsets choisis par l'utilisateur parmi l'ensemble I d'items. L'utilisateur dispose de 8 opérateurs de manipulation déclenchés par un clic souris sur un sommet. En supposant que le sommet sélection soit X voici les deux opérateurs principaux :

- O_1, *spécialisation*. Cet opérateur développe tous les arcs sortants de X. C'est-à-dire, les règles *simples* dont la prémisse est l'itemset X, $O_1(X) = \{X \rightarrow i \,|\, i \in I \setminus X \, et \, C(X \rightarrow i)\}$
- O_2, *généralisation*. Cet opérateur symétrique développe tous les arc entrants de X. Il s'agit de toutes les règles *simples* $O_2(X) = \{X \setminus \{i\} \rightarrow X \,|\, i \in X\}$ vérifiant le critère C.
- La figure 5.3 illustre l'opérateur O_1.

Ce choix pour les opérateurs est fondé sur l'étude des processus de raisonnement menée dans [117], qui souligne que Spécialisation et généralisation sont les deux processus cognitifs fondamentaux pour la génération de nouvelles règles. Ces deux opérateurs basiques permettent d'en dériver 6 autres

selon qu'on les étend aux règles *non-simples*, ou qu'on les rend *récursifs* par application successive sur tous les sommets nouvellement produits jusqu'à stabilité. Il est important de noter que d'une part ces opérateurs ont une sémantique claire pour l'utilisateur et d'autre part leur complexité algorithmique en faible (linéaire avec le nombre d'attributs $|I|$). Dans [146], nous présentons de manière détaillée les opérateurs d'interaction et les algorithmes sous-jacents.

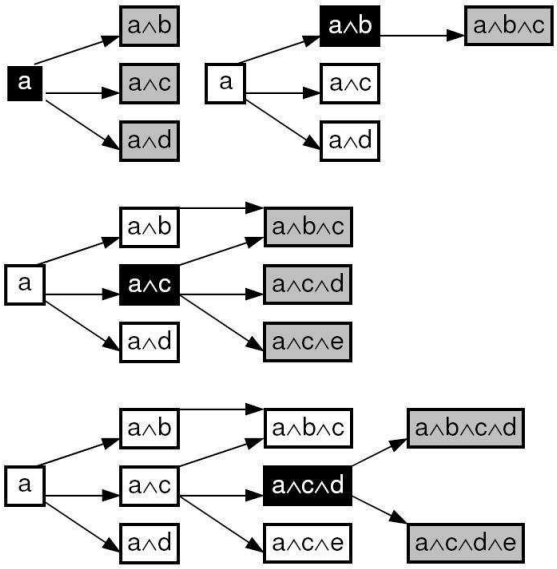


FIG. 5.3 – Des explorations locales successives dans l'ensemble $\mathcal{R}$ des règles

5.3.2.2 Visualisation dynamique

Les opérateurs d'interaction imposent la gestion de la dynamique de la visualisation. Afin de préserver la carte mentale de l'utilisateur, les écarts entre deux graphes successifs doivent être minimisés. A cet effet, deux contraintes de lisibilité ont été intégrées dans le placement des sommets du graphe :

- Une contrainte *esthétique* : minimiser le nombre d'interaction et la longeur des arcs ([61, 187])
- Une contrainte de *stabilisation* : minimiser la différence entre deux graphes successifs (cf [106] et [147])

Ces deux contraintes étant antagoniste, un compromis est calculé à l'aide d'un algorithme génétique spécifique (cf [106] [146]). Une expérimentation de cette approche sur un corpus issu d'une étude marketing est détaillée dans [146].

5.4 Visualisation en 3D et Réalité Virtuelle

5.4.1 Représentation en 3D

Afin de supporter de grandes quantités de règles tout en mettant en évidence les meilleures d'entre elles, nous avons choisi une représentation 3D pour implémenter la méthodologie *RF*. Plus précisément, la visualisation de chaque sous-ensemble de règles repose sur la *métaphore* du paysage d'information (*information landscape* [7]). Ce paysage 3D est constitué d'une *arène* semi-circulaire (figure 5.4), qui sert de repère à l'utilisateur et contraint le placement des règles. Chaque règle y est graphiquement représentée par une *sphère* posée sur un *cône*, et une étiquette textuelle complémentaire rappelle ses caractéristiques (figure 5.5).

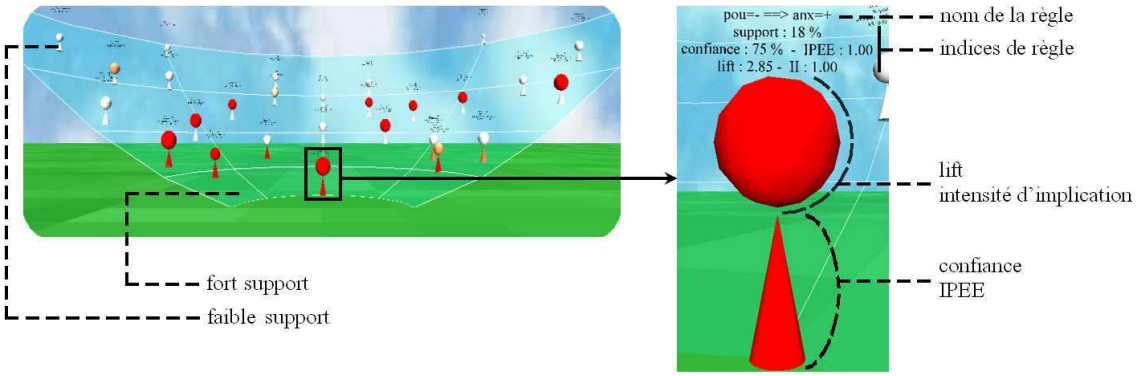


FIG. 5.5 – Encodage graphique dans *ARVis 1.2*

Nous optons également pour une intégration des mesures de qualité dans la représentation graphique en considérant les critères psychovisuels évalués par [15] [213]L'encodage graphique choisi pour les mesures de qualité est le suivant :

- la *position* de l'objet représente le *support*,
- la *surface* visible du *cône* représente la *confiance*,
- la *surface* visible de la *sphère* représente le *lift* (normalisé entre 0 et 1),
- la *luminosité* de la *sphère* représente l'*intensité d'implication*,
- la *luminosité* du *cône* représente *IPEE*.

Ainsi, L'encodage graphique est conçu afin de mettre en évidence les règles de bonne qualité (proches, volumineuses, et lumineuses). De plus, la *sphère* est dédiée aux mesures d'écart à l'*indépendance*, tandis que le *cône* est dédié aux indices d'écart à l'*équilibre*.

Deux métaphores graphiques différentes ont été implémentées en VRML dans les versions 1.1 et 1.2 de l'outil ARVis (Association Rule Visualisation).

La description précédente correspond à la version *ARVis 1.2*. La première version, *ARVis 1.1*, utilise un encodage graphique différente pour les mesures de qualité, et le paysage est constitué de deux arènes en vis à vis : une pour les règles spécifiques et une pour les règles générales. Les règles de l'arène spécifique ne supportent que l'opérateur de spécialisation, et réciproquement l'arène générale l'opérateur de généralisation. De plus, afin d'anticiper l'effet des opérateurs, leur résultat est affiché en miniature au sein la sphère correspondante (figure 5.6).

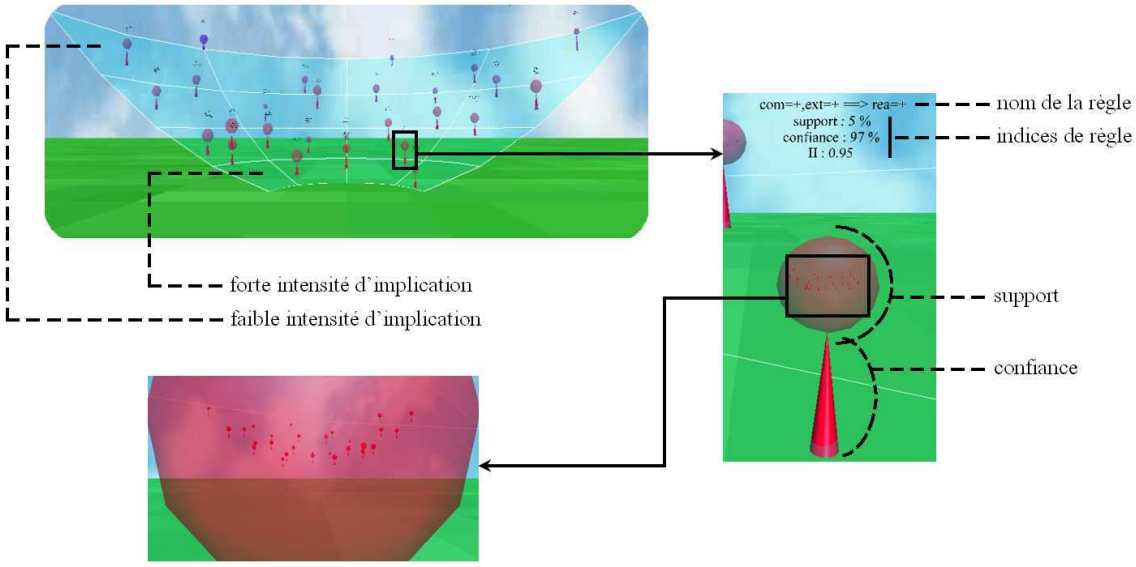


FIG. 5.6 – Encodage graphique dans *ARVis 1.1*

5.4.2 Opérateurs d'interaction

Huit opérateurs d'interaction (appelés relations de voisinage), étendant les opérateurs de généralisation et de spécialisation définis dans Felix (section 5.3.2.1), sont proposés dans *ARVis 1.2*. Ils sont tous associés aux mesures de qualité à travers le critère de significativité $C(r)$ d'une règle r de la manière suivant : considérant l'ensemble de mesures $M = sp, cf, li, ii, ip$ (i.e. support, confiance, lift, intensité d'implication et IPEE), dont chaque mesure m est associée à l'intervalle $[min_m,\ max_m]$ choisi par l'utilisateur, une règle r est conservée ssi : $C(r)$ est vérifiée, avec $C(r) = \forall m \in M, min_m \leq m(r) \leq max_m$. Les seuils peuvent être modifiés par l'utilisateur à tout instant pendant la navigation.

Partant des deux opérateurs basiques Π_1/Π_2 de *spécialisation/généralisation* (cf section 5.3.2.1), six nouveaux opérateurs sont dérivés :

1. *Spécialisation concordante* :

$$\Pi_3(X \rightarrow y) = \{X \cup \{z\} \rightarrow y \mid z \in I \setminus (X \cup y) \; et \; C(X \cup \{z\} \rightarrow y)\}$$

2. *Spécialisation d'exception* :

$$\Pi_4(X \rightarrow y) = \{X \cup \{z\} \rightarrow \overline{y} \mid z \in I \setminus (X \cup \{\overline{y}\}) \; et \; C(X \cup \{z\} \rightarrow \overline{y})\}$$

où $\overline{y}$ désigne n'importe quel item provenant de la même variable que y mais présentant une modalité différente.

3. *Généralisation de la prémisse* :

$$\Pi_5(X \rightarrow y) = \{X \setminus z \rightarrow z \mid z \in X \; et \; C(X \setminus z \rightarrow z)\}$$

4. *Prémisse commune* :

$$\Pi_6(X \rightarrow y) = \{X \rightarrow z \mid z \in I \setminus X \; et \; C(X \rightarrow z)\}$$

5. *Conclusion commune* :

$$\Pi_7(X \rightarrow y) = \{z \rightarrow y \mid z \in I \setminus \{y\} \; et \; C(z \rightarrow y)\}$$

6. *Items communs* :

$$\Pi_8(X \rightarrow y) = \{(X \cup \{y\}) \setminus \{z\} \rightarrow z \mid z \in X \cup \{y\} \; et \; C((X \cup \{y\}) \setminus \{z\} \rightarrow z)\}$$

Dans [117], Holland *et al.* soulignent qu'une règle trop générale peut être spécialisée en deux types de règles complémentaires : les règles exceptions et les règles concordantes. Les règles exceptions (voir par exemple [118] et [206]) visent à expliquer les contre-exemples de la règle générale, tandis que les règles concordantes visent à mieux expliquer les exemples.

La relation *Généralisation* consiste à simplifier la prémisse d'une règle (processus de simplification des conditions décrit dans [116]). Elle est complémentaire à *Spécialisation concordante*, puisque après avoir appliqué *Spécialisation concordante* sur une règle r, on peut retrouver r en utilisant *Généralisation*. La relation *Généralisation de la prémisse* est quant à elle reprise de *ARVis 1.1*.

Les relations *Prémisse commune* et *Conclusion commune* préservent la prémisse et changent la conclusion, ou vice versa. *Items communs* permet de permuter les items dans une règle. Toutes les règles produites par cette relation sont vérifiées par la même population d'individus dans les données.

Imaginons que l'utilisateur applique une relation de voisinage Π sur une règle r. Ceci génère et affiche un nouveau sous-ensemble $S = \Pi(r)$ contenant toutes les règles voisines de r selon Π. Nous appelons r la *règle de transition*, car c'est elle qui permet le passage d'un sous-ensemble à un autre. En fonction de la relation Π choisie, S peut contenir ou ne pas contenir la règle de transition (les relations de voisinage ne sont pas nécessairement réflexives). Dans *ARVis 1.2*, nous ajoutons systématiquement la règle de transition à tout sous-ensemble généré par une relation de voisinage. Ceci permet de pouvoir effectuer des comparaisons entre la règle de transition et ses règles voisines. Par exemple, avec la relation de voisinage *Généralisation*, il est intéressant de comparer une règle à ses voisines afin de repérer les items superflus dans la règle (ceux dont la suppression ne dégrade pas la qualité de la règle). Réciproquement, avec la relation *Spécialisation concordante*, comparer une règle à ses voisines permet de vérifier si l'ajout d'un nouvel item en prémisse améliore ou non la prédiction de la conclusion.

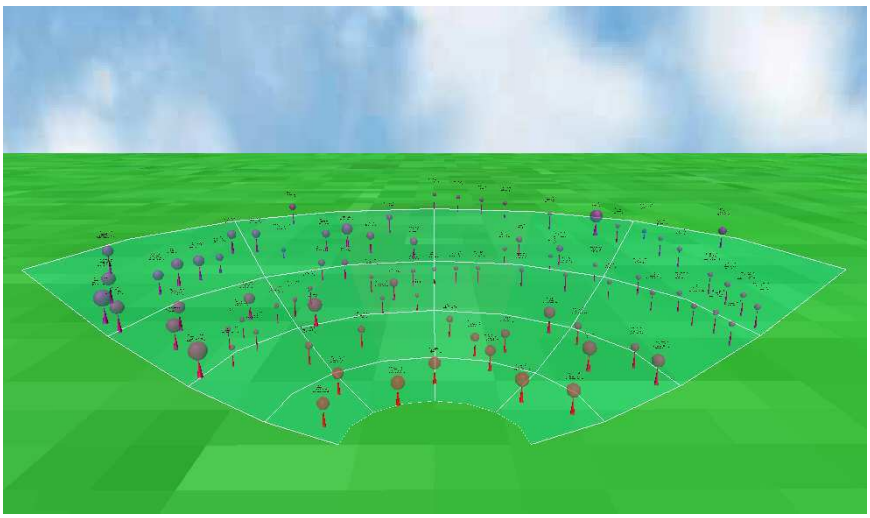
5.5 Synthèse des publications

Partant de la problématique de post-traitement des règles d'association, nous avons dans un premier temps proposé une première tentative de visualisation par des graphes de règles, détaillée dans [88]. Dans ses travaux de thèse [152], Rémi Lehn a ensuite poursuivi dans cette direction, et proposé une première approche de fouille interactive de règles fondée sur la visualisation par des graphes [145]. Cette approche a été implémentée dans l'outil Felix, présenté dans [156] et [155]. Un algorithme génétique spécifique à l'optimisation des contraintes de tracé induites par le graphe est détaillé dans [157] et [148] et affiné dans [147] [149]. Une synthèse de ce travail est publiée dans la revue InCognito [146].

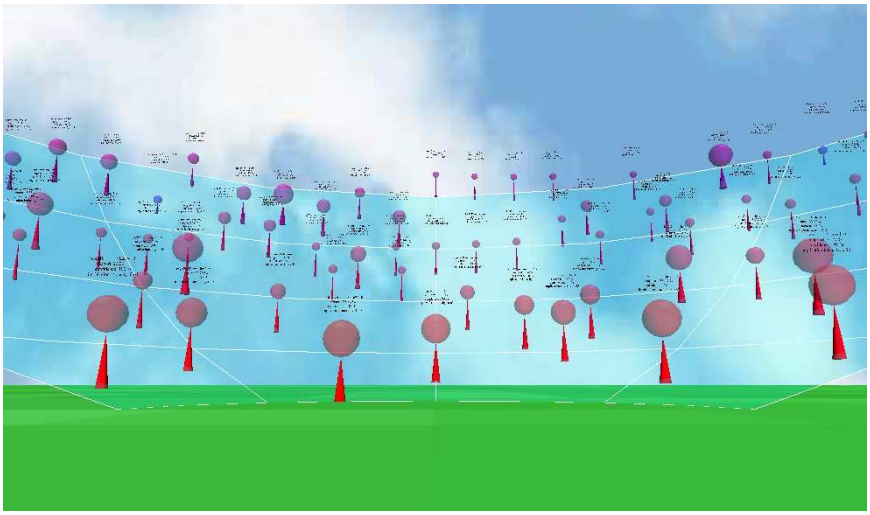
Dans la poursuite de ces travaux sur la visualisation par des graphes, une seconde approche fondée sur une visualisation 3D et la réalité virtuelle a été traitée dans la thèse de Julien Blanchard [16] (cf section 5.4). Les principes de cette approche sont présentés dans [20]. L'intégration des mesures de qualité a été précisée dans [19] [18] [28], et les aspects réalité virtuelle dans [33] et[32]. Une synthèse de ce travail est publiée dans les revues InCognito[21] et KAIS[22].

En synthèse, ces approches antropocentrées pour la fouille de règles d'association disposent de trois avantages majeurs : ils offrent un véritable support visuel interactif pour l'aide à la décision sur les règles d'association, ils permettent la prise en compte des critères subjectifs relatifs aux préférences de l'utilisateur, et incorporent des algorithmes d'extraction locale qui

permettent de conserver des temps de calcul acceptables lors du passage à l'échelle sur des données volumineuses.



(a)



(b)

(c)

FIG. 5.4 – Un paysage de règles dans *ARVis*

Chapitre 6

Gestion de connaissances

Sommaire

Parallèlement, stimulés par des projets émanant d'entreprises, et dans une perspective initiale plus applicative, nous nous sommes intéressés à la gestion et l'ingénierie des connaissances [68] [63], et à son prolongement récent dans le cadre du web sémantique. Ce chapitre résume partiellement nos principaux apports scientifiques et développements de logiciels dans ce domaine. Afin de mieux appréhender nos travaux, *trois articles complémentaires joints en annexe* détaillent et complètent les trois principales sections de ce chapitre.

Dans un premier temps, nous présentons l'approche « serveur de connaissances » que nous avons conçue et implémentée dans deux outils SAMANTA

puis Athanor, en partenariat avec La Poste puis Performance SA. L'article fournit en annexe C vient compléter cette partie.

Puis, nous résumons nos travaux sur la modélisation de connaissances pour des agents émotionnels. Ce travail, dont l'objectif in fine est de produire un système d'aide à la décision sur la dynamique des groupes, a été impulsé par le projet ARTA sur l'étude du comportement en milieu professionnel des victimes d'un traumatisme crânien. Il s'est ensuite prolongé dans le cadre d'un projet RIAM : Groupe d'Agents Collaboratifs Emotionnels (GRACE). Cette section détaille le modèle émotionnel EST que nous avons proposé, ainsi que les modélisations UML/AUML utilisées. Cette synthèse est complétée par l'article fourni en annexe D.

Enfin, la dernière section synthétise nos travaux récents sur l'alignement d'ontologies. L'approche proposée a l'originalité d'être extensionnelle et asymétrique, et de coupler la découverte de règles d'association en fouille de données à l'alignement d'ontologies. Après avoir décrit notre approche implémentée dans l'outil Aroma, nous présentons très succinctement les résultats des expériences menées sur différents bancs d'essai. L'article fourni en annexe E illustre et complète cette dernière partie.

6.1 Approche Serveur de connaissances

Partant d'une problématique de maintenance et de diagnostic de pannes sur les machines de tri postal, menée en collaboration avec La Poste, nous avons conçu une approche « serveur de connaissances » répondant à leurs besoins en gestion des connaissances. L'idée sous-jacente à cette approche « serveur de connaissances », était de concevoir un serveur d'un nouveau type, analogue à un serveur web mais transposant ses services à la connaissance. Ce type de serveur bénéficie ainsi des technologies du web, simplifiant son déploiement, facilitant l'intégration de ressources multiples (connaissances, documentation, plan, images, fiches d'intervention, nomenclatures, ...), et offrant des capacité de traitement des connaissances.

Parallèlement à une campagne de recueil de connaissances par entretiens avec les experts de ces machines, nous avons implémenté notre approche dans l'outil SAMANTA (Système d'Aide à la MAintenance des Trieuses Automatiques). Ce logiciel est fondé 3 ontologies complémentaires : une ontologie pour les tâches de diagnostic, une ontologie décrivant la composition d'une machine de tri, et une ontologie des symptômes. L'outil incorpore un éditeur d'ontologies entièrement graphique garantissant l'évolutivité du système, permet l'association à des ressources multimédia complémentaires (dont des modèles de machine de tri en réalité virtuelle), et enfin intègre un raisonneur

prolog permettant d'offrir notamment une fonctionnalité d'aide à la décision pour le diagnostic.

Nous avons ensuite conçu Athanor, une généralisation de cette approche serveur de connaissances, incorporant une quatrième ontologie pour modéliser les compétences, et étendu l'ontologie des tâches de diagnostic à des tâches quelconques. Ce second logiciel a été réalisé dans le cadre d'une collaboration avec la société performanSE SA.

6.1.1 Gestion des connaissances et mémoires organisationnelles

La gestion des connaissances constitue l'un des problèmes stratégiques auxquels sont confrontées les entreprises. Cette problématique est amplifiée par trois phénomènes cumulatifs : d'une part l'accroissement des données et des documents stockés stimulé par l'usage des environnements de travail collaboratif, les collecticiels ; et d'autre part la baisse du niveau d'expertise dans certains domaines causée par l'érosion du nombre d'experts (pyramide des âges qui vieillit, mobilité importante), et enfin la réduction du cycle de vie des produits imposée un contexte fortement concurrentiel.

Dans ce contexte, la gestion des connaissances devient nécessaire pour assurer la conservation des connaissances acquises. Cette mémorisation structurée des connaissances est d'autant plus difficile à effectuer que d'une part, les structures des produits sont de plus en plus complexes (multiplicité des technologies et des acteurs, grand nombre de composants), et que d'autre part les caractéristiques des sources de connaissances relatives à ces systèmes sont de multiples sortes [6] [66] (individuelles, collectives, tacites, procédurales, ...).

La Gestion des Connaissances (GC) peut être définie comme une approche qui regroupe l'ensemble des méthodes et techniques permettant de formaliser, partager, diffuser et enrichir le capital intellectuel de l'entreprise. L'objectif sous-jacent est de pouvoir fournir à chaque classe d'utilisateurs les connaissances pertinentes au bon moment. Il faut noter que l'utilisation de l'informatique n'est a priori pas obligatoire en GC. Mais, comme le font remarquer de nombreux d'auteurs [6] [64] [178], cette dernière est amenée à jouer un rôle croissant. Au-delà des fonctionnalités qu'elle apporte (e.g. diffusion de l'information, utilisation d'éditeurs graphiques adaptés), l'utilisation de l'informatique permet de renforcer l'interactivité avec l'utilisateur en opérant un glissement vers les systèmes antropocentrés et l'aide à la décision.

L'approche serveur de connaissances a pour objet la gestion d'une mémoire d'entreprise. Le processus de création d'une telle mémoire est considéré comme le passage d'une mémoire de travail à une mémoire organisationnelle.

Elle se définit comme un capital de connaissances accessible indépendamment des acteurs qui l'ont créée [186].

Une démarche de gestion des connaissances est associée aux différentes phases du cycle de vie d'une mémoire d'entreprise [64] : de la détection des besoins jusqu'à son utilisation. Elle peut également s'appuyer sur les méthodes de formalisation conceptuelle en ingénierie des connaissances. Citons par exemple la méthode KADS [196] de conception de systèmes experts, ou encore la méthode de décomposition systémique et sémiotique MKSM [68] qui amène à la construction d'un livre connaissances. Le degré de formalisation des connaissances (ontologies, modélisation objets, réseaux sémantiques, règles de productions, ...) doit être d'autant plus précis que l'on veut assurer un niveau de traitement sémantique élevé.

Ensuite, une fois recensées les sources de connaissances disponibles et valides qui peuvent être utilisées (documentations papiers, experts humains, bases de données, ...), plusieurs types de mémoires sont envisageables : e.g. une Gestion Electronique de Documents (GED), une mémoire à base de cas, une documentation technique intelligente [181], une mémoire à base de graphes conceptuels (Corese [45, 46]), ou encore, comme ici, un serveur de connaissances. Enfin, plus récemment, les modèles et outils développés dans le cadre du web sémantique [49], ouvrent de nouvelles perspectives pour le développement de systèmes de gestion des connaissances standardisés et interopérables.

6.1.2 Serveur de connaissances Athanor

6.1.2.1 Les modèles de connaissances d'ATHANOR

Le serveur de connaissances ATHANOR est fondé sur trois modèles ontologiques principaux permettant de décrire les connaissances selon trois points de vue complémentaires : un modèle processus (ou tâche) orienté vers les processus métiers (comment), un modèle organique du système supportant les processus (quoi), et un double modèle compétences/organigramme (qui). A un niveau plus global, ces trois modèles entretiennent de fortes interrelations afin d'exprimer la richesse multidimensionnelle des connaissances : les connaissances concernent des processus portant sur des composants d'un système dont la manipulation nécessite des compétences mises en oeuvre par des individus (Fig 6.1).

Plus précisément, le modèle processus a pour objet la description de processus métier pratiqués par des experts. Il permet de maintenir des connaissances de nature procédurale qui sont décomposées en un enchaînement d'étapes. Chaque étape y est décrite en fonction de ses propriétés, et l'enchaî-

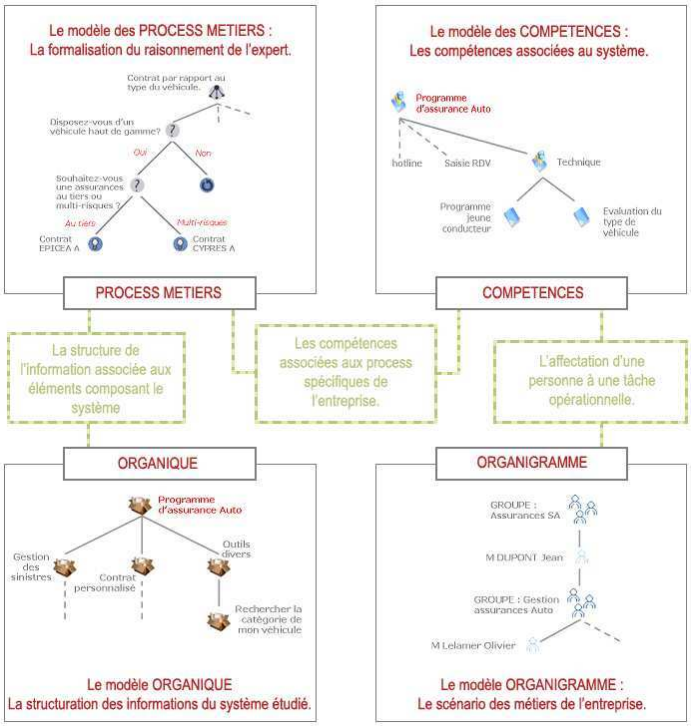


FIG. 6.1 – Les modèles ontologiques et leurs interactions

nement des étapes traduit des règles de raisonnement. Le modèle organique permet de décrire la structure du système à travers sa décomposition en une hiérarchie de composants et de groupe de composants. Enfin, le double modèle compétences/organigramme permet de décrire d'une part une hiérarchie de compétences requises pour mener les étapes du processus, et d'autre part la hiérarchie des acteurs et de leurs compétences acquises sur le système. Chaque compétence est définie en terme de savoir, savoir-faire et savoir-être.

Dans un souci de simplification de l'accès à cet ensemble de modèles, ceux-ci sont conçus afin de disposer d'une représentation graphique canonique : un arbre ou un graphe.

6.1.2.2 Architecture du serveur de connaissances

D'un point de vue plus technique, le serveur de connaissances ATHANOR est conçu de manière modulaire et extensible, et s'appuie sur un support technique universel, Internet, facilitant le déploiement des connaissances. Plus précisément, l'architecture technique comporte en son coeur un serveur de connaissances gérant une base de connaissances, et en périphérie un ensemble

de modules graphiques proposant différentes vues fonctionnelles sur les modèles (cf. fig. 6.2) :

- Un module *Praticien* pour activer les connaissances en mode résolution de problème,
- Un module *Expert* pour décrire et mettre à jour les modèles et leurs associations,
- Un module *Manager* pour surveiller avec un tableau de bord la base de connaissance et son utilisation,
- Un module *Pédagogue* pour consulter les connaissances sous forme graphique commentée.

En complément, un ensemble de modules additionnels sont proposés pour faciliter à la fois l'intégration d'ATHANOR dans un système d'information préexistant et y promouvoir son utilisation :

- Un module *réalité virtuelle* en liaison avec le modèle organique, pour l'utilisation de représentations en trois dimensions du système supportant les connaissances,
- Un module de *communication*, qui permet aux utilisateurs s'échanger des informations,
- Un module de *gestion de groupes*, qui permet gérer les utilisateurs et leurs droits d'accès,
- Un module d'*expertise nomade*, générant une version allégée des modèles pour des périphériques tel qu'un PDA,
- Un module d*'accès au système de gestion documentaire du SI*, permettant le référencement de documents externes dans la base de connaissances.

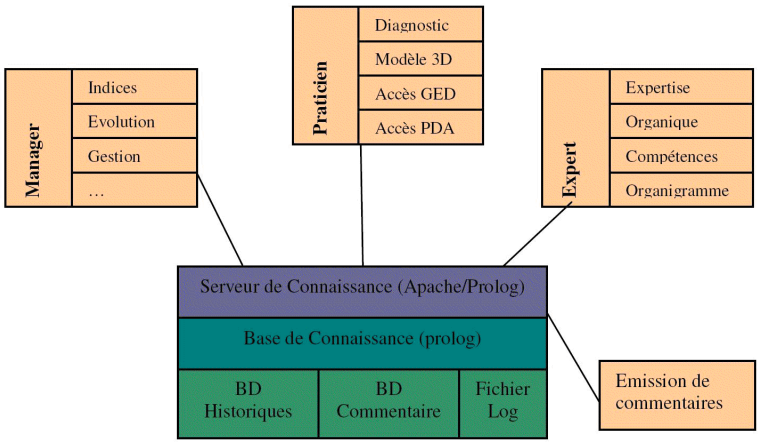


FIG. 6.2 – Architecture Modulaire du Serveur de Connaissances

La gestion de la complexité des systèmes sur lesquels porte les connaissances est facilitée par l'utilisation des représentations graphiques associées

aux trois modèles, afin d'assurer la lisibilité et l'appropriation des connaissances. Les modules Expert et Pédagogue permettent d'organiser les connaissances procédurales, d'appréhender la complexité structurelle du système, et de gérer les compétences associées à ces modèles. Le module Manager, sorte de tableau de bord sur la base de connaissances, a deux objectifs : l'un est d'offrir une synthèse sur la structure de la base elle-même, c'est l'étude de la structure *statique* des connaissances ; et l'autre, considère son aspect *dynamique*, la manière dont elle évolue au fur et à mesure des utilisations. Le module Praticien est accessible à tous les utilisateurs. Il permet d'actionner la base de connaissances, en déroulant les processus métiers par un jeu de questionnement, et suivant une problématique de résolution de problème.

Chacun de ces modules possède sa propre interface graphique avec l'utilisateur et est conçu comme un client séparé fonctionnant dans un navigateur web, s'appuyant sur les technologies de l'Internet (dhtml, java, ...) et communiquant par requêtes avec le noyau Apache/Prolog du serveur de connaissances. En interne, le serveur de connaissances maintient les modèles de connaissances dans des bases de connaissances opérationnelles implémentées en prolog. Cette implémentation en prolog est transparente pour les utilisateurs et est totalement cachée par les interfaces graphiques des différents modules. Ce choix du langage Prolog pour l'implémentation du serveur a l'avantage de faciliter la gestion interne des connaissances recueillies, mais surtout de réaliser un stockage opérationnel des connaissances permettant leur activation à travers des moteurs d'inférence en mode résolution de problèmes. Il offre également de grandes perspectives d'extension du modèle actuel de connaissances.

Enfin, en réponse au besoin de *diffusion des connaissances*, le choix Intranet/Apache facilite l'accès au serveur de connaissances sur un réseau d'entreprise. Il permettra aussi d'en observer l'utilisation par analyse des fichiers log, à l'aide de techniques de découverte de connaissances.

6.1.2.3 Exemple de Formalisation Graphique des Connaissances : le modèle processus

Ainsi que les modèles *organique* et *compétences/organigramme*, le modèle *processus* dispose d'une représentation graphique canonique : un graphe. Ce modèle permet de représenter des connaissances procédurales liées à un savoir-faire sous la forme d'un processus décomposé en une suite d'étapes à mener. Ainsi, dans le cas d'un diagnostic, chaque étape consiste à tester des hypothèses sur l'état des composants ou des fonctionnalités du système. L'enchaînement des étapes suit une logique d'efficacité : des hypothèses les plus simples aux plus complexes.

Le graphe du processus est appelé *logigramme* (Fig 6.3), et le raisonnement de l'expert s'y traduit par un parcours du graphe depuis sa racine (de haut en bas et de gauche à droite). Chaque étape du raisonnement correspondant à un sommet de ce graphe. Cette représentation graphique peut être vue comme une généralisation des arbres de décision et des arbres de défaillance, les enrichissant de sommets structurants afin de prendre en compte le modèle de raisonnement des experts.

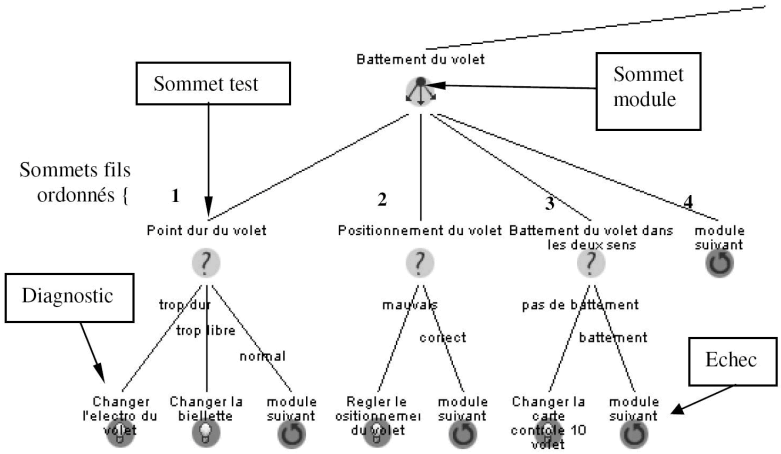


FIG. 6.3 – Représentation par logigramme du modèle processus

En complément de cette logique d'enchaînement portée par la structure globale du logigramme, nous avons défini quatre types de sommets dont la sémantique diffère. Les deux premiers types sont des sommets intermédiaires structurants :

- Les sommets *test* associés à une variable, typiques des arbres de décision, dont les fils ne sont pas ordonnés, mais dont chaque arc est associé à une valeur de la variable. La variable est généralement associée à l'état de fonctionnement dans lequel se trouve un élément du système à diagnostiquer.

- Les sommets *module*, absents des arbres de décision et de défaillance, dont les fils sont ordonnés de gauche à droite et généralement du plus simple au plus complexe, au sens de l'expert. Chacun de ces sommets permet de définir un module de connaissances.

Les sommets *module* permettent ainsi d'intégrer des principes cognitifs caractéristiques des stratégies de décision expertes [10], dont un principe de parcimonie/décidabilité :

- Les premiers sommets fils d'un module permettent d'arriver à une décision à moindre coût par des opérations simples (parcimonie)

- Les sommets fils suivant offrent la possibilité de réaliser des opérations de plus en plus complexe afin d'arriver à une prise de décision même si elle s'avère coûteuse (décidabilité).

Nous proposons aussi deux autres types pour les sommets terminaux (feuilles) :

- Les sommets *diagnostic*, indiquant la fin du processus : résolution du problème et la réparation à effectuer.

- Les sommets associés à un *échec* provoquant la mise en oeuvre d'un mécanisme de retour par back-tracking au dernier sommet module traité et la transition au sommet suivant au sens de l'ordre induit par ce sommet module.

Enfin, chaque sommet cette représentation est décrit par des informations propres :

(a) le composant suspecté : élément supposé défectueux à ce stade du raisonnement (composant décrit dans le modèle organique)

(b) L'accès a une représentation en réalité virtuelle du composant suspecté afin de faciliter sa localisation (associé au modèle organique).

(c) La description multimédia du mode opératoire à suivre pouvant inclure des liaisons vers une documentation complémentaire, des images, des vidéos, des sons, des modèles en réalité virtuelle, des odeurs, ainsi que des informations connexes du type instrument à utiliser pour évaluer l'état du ou des composants en cause.

(d) La liste des compétences requises pour opérer cette intervention (associée au modèle compétences) et les personnes de l'organisation qui possèdent ces compétences.

Une propriété importante de cette formalisation graphique, réside dans la possibilité de la transformer en un ensemble de règles de production, en traduisant l'ensemble des chemins menant de la racine à chacune des feuilles. Cette représentation graphique des connaissances procédurales a l'avantage d'être beaucoup plus intelligible et synthétique qu'un ensemble équivalent de règles de production.

Les graphes utilisés pour la représentation graphique des ontologies ont également fait l'objet d'une optimisation de leur placement, afin d'en améliorer la lisibilité [185, 184].

6.2 Modélisation de connaissances émotionnelles

En collaboration avec la société PerformanSE SA et l'association APARTA [1], nous nous sommes intéressés à la modélisation de comportements de groupes appliquée à la réinsertion professionnelle des traumatisés crâniens. Plus précisément, nous avons réalisé une modélisation de leur comportement afin de le simuler dans un environnement multi-agents. Les spécificités de la population étudiée, nous ont amenés à définir un modèle particulier d'agent, accordant un accent à la dimension émotionnelle.

Dans le domaine des systèmes multi-agents [215], les travaux de recherche traitant d'agents émotionnels sont basés sur le modèle cognitif OCC (Ortony et al. 1988) [177] et portent sur l'aspect social des émotions ("Affective Reasoner" (Kapoor et al. 2001) [137], PETEEI (El-Nasr et al. 2000) [67]), ou sur le lien entre planification et émotions (MRE (Rickel et al. 2002) [190]). Toutefois, les notions de croyance et de désir ne sont pas prises en compte, ce qui les éloigne du comportement humain proposé par (Bratman 1987) [36].

En nous appuyant sur le modèle BDE (Florea et al. 2003) [72], et en considérant des agents de type BDI (Rao et al. 1991) [189] dotés d'émotions basées sur OCC, nous avons proposé d'étendre la modélisation des émotions au sein des agents. Pour cela, en liaison avec des psychologues, nous avons décliné les émotions d'une part à travers la *personnalité* d'un agent, et d'autre part à travers les *interactions* entre agents. Nous avons également modélisé l'ensemble de l'expertise recueillie à l'aide d'UML et d'AgentUML[176]. Puis, les modèles obtenus ont été implémentés dans la plateforme JADE.

6.2.1 Modélisation EST de la personnalité d'un agent émotionnel

L'ensemble du comportement d'un agent est déjà bien défini dans les modèles BDI et OCC. Les agents disposent en effet d'une mémoire, et d'un système procédural qui régit leurs activités en fonction de leurs buts et de leur perception de l'environnement. Notre première extension réside dans la modélisation de la personnalité d'un agent. Afin de mieux définir la dimension émotionnelle, nous avons défini un modèle EST de la personnalité. Le modèle EST est détaillé dans la figure 6.4. Il structuré en trois niveaux principaux, les *émotions*, les *sentiments* et le *tempérament*, régis par des vitesses d'évolution décroissantes ; et est complété par une description des états psychologiques et physiques.

Les variables mémorisées dans ce modèle sont les suivantes :

[1] Atelier Protégé d'Aide à la Réinsertion des Traumatisés crâniens Atlantique

- Les *émotions* sont celles du modèle OCC : 11 couples d'émotions antagonistes : *joie / tristesse* (JOI/TRI), *pitié / joie malsaine* (PIT/JOM), *content pour / jalousie* (CON/JAL), *espoir / peur* (ESP/PEU), *soulagement / déception* (SOU/DEC), *satisfaction / horreur* (SAT/HOR), *fierté / honte* (FRT/HON), *engouement / indignation* (ENG/IND), *gratitude / colère* (GRA/COL), *gratification / remord* (GRF/REM), *goût / dégoût* (GOU/DEG). L'évolution des émotions est déclenchée par l'activité de l'agent et est modulée par son tempérament. Par exemple, une extraversion élevée renforcera l'émotion de colère.
- Les *sentiments* historisent les émotions et constituent une sorte de mémoire émotionnelle. Elles sont au nombre de 3 : *quiétude*, *amitié*, *confiance*. L'évolution des sentiments est conditionnée par l'intensité des émotions ressenties sur une période de temps.
- Le *tempérament* d'une personne, considéré comme figé, est basé sur le modèle PerformanSe (Gras et al. 2003) [92] selon dix traits de personnalité : *extraversion / introversion* (EXT/INT), *détente / anxiété* (DET/ANX), *combativité / conciliation* (COM/CON), *réceptivité / détermination* (REC/DET), *affirmation / remise en cause* (AFF/REC), *dynamisme / conformisme intellectuel* (DIN/CIN), *rigueur / improvisation* (RIG/IMP), *réalisation / facilitation* (REA/FAC), *appartenance / indépendance* (APP/IND), *pouvoir / protection* (POU/PRO).
- En complément, l'agent dispose d'un *état psychologique* qui traduit son état mental vis-à-vis de son travail, et un *état physique*.

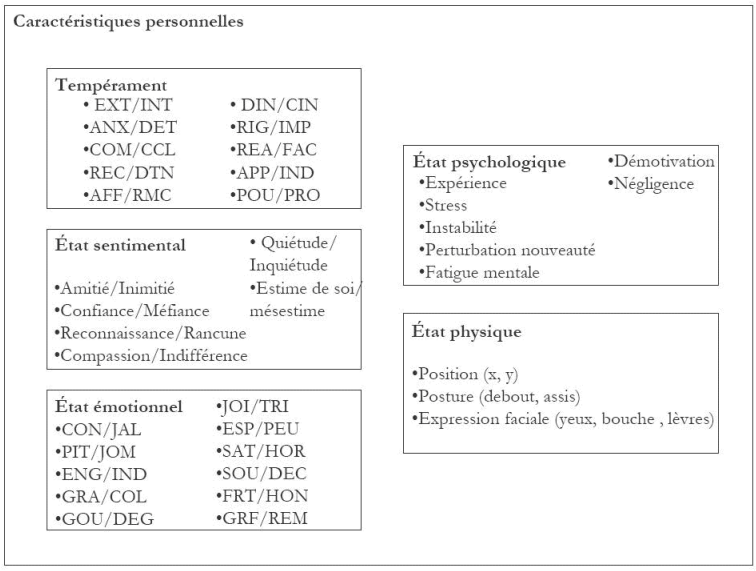


FIG. 6.4 – Modèle émotionnel EST

6.2.2 Modélisation UML d'un agent émotionnel

La modélisation psychologique EST d'un agent a été traduite en UML. La figure 6.5 décrit le diagramme de classes d'un agent émotionnel. On y trouve le modèle EST et ses relations avec le modèle de sa mémoire, du fonctionnement BDI, et du mécanisme d'évolution de l'état de l'agent. La figure 6.6 détaille le diagramme d'activité de l'agent, qui décrit son évolution conformément aux principes du modèle BDI.

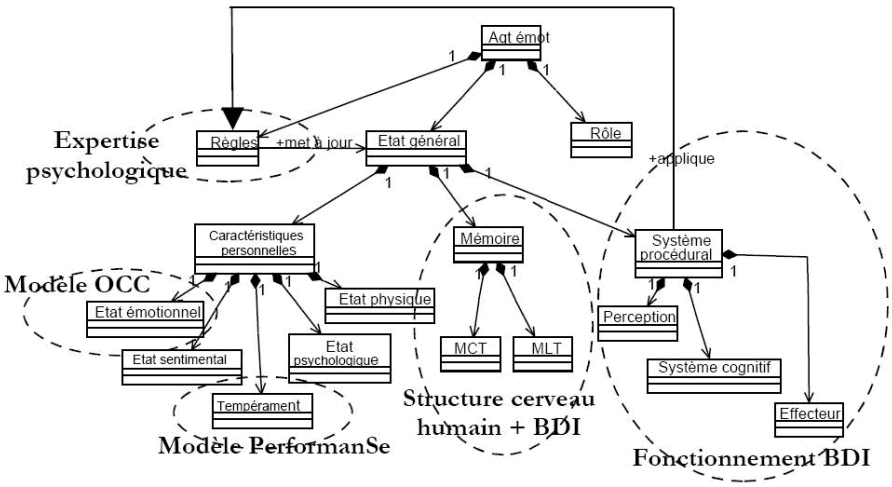


FIG. 6.5 – Modélisation UML d'un agent émotionnel

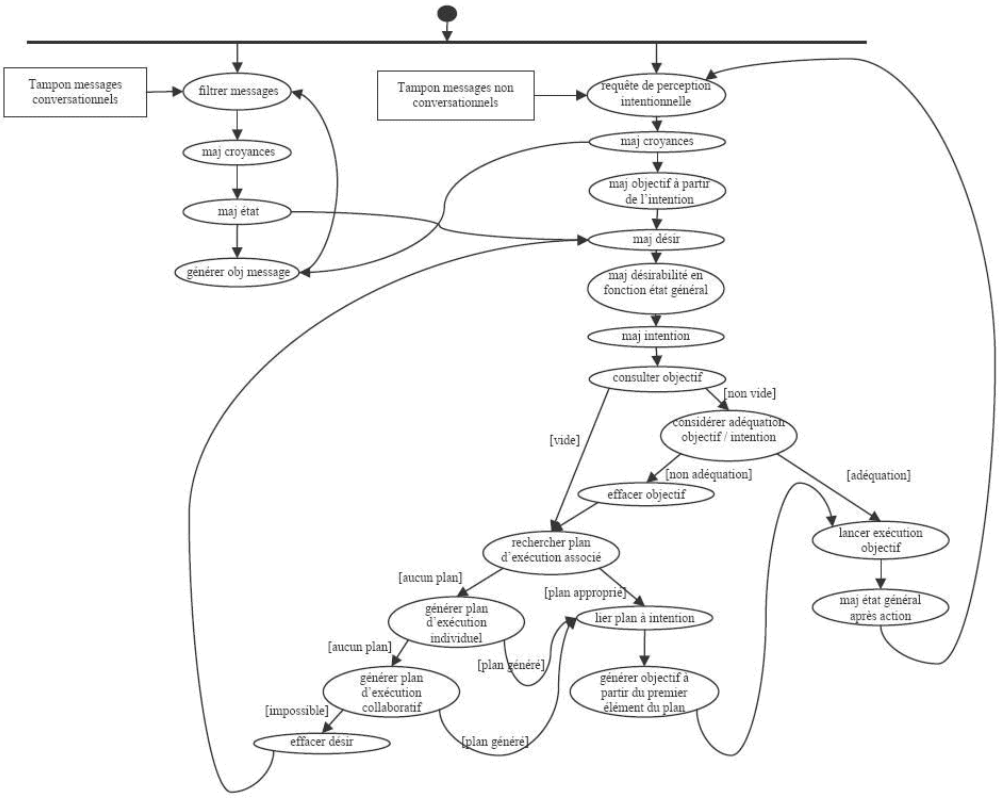


FIG. 6.6 – Diagramme UML d'activité d'un agent

6.2.3 Modélisation des interactions

L'ensemble des interactions entre agents sur chaque poste de la chaîne de production a fait l'objet d'une modélisation AgentUML [176]. Notamment les diagrammes d'activité et de séquence nous ont permis de décrire la dynamique des interactions entre agent en fonction des postes occupés. Les figures 6.7 et 6.8 illustrent les interactions entre agents sur le poste 7. En particulier, le modèle de séquence 6.8 permet de décrire les règles d'interaction entre agents en incorporant leurs caractéristiques émotionnelles.

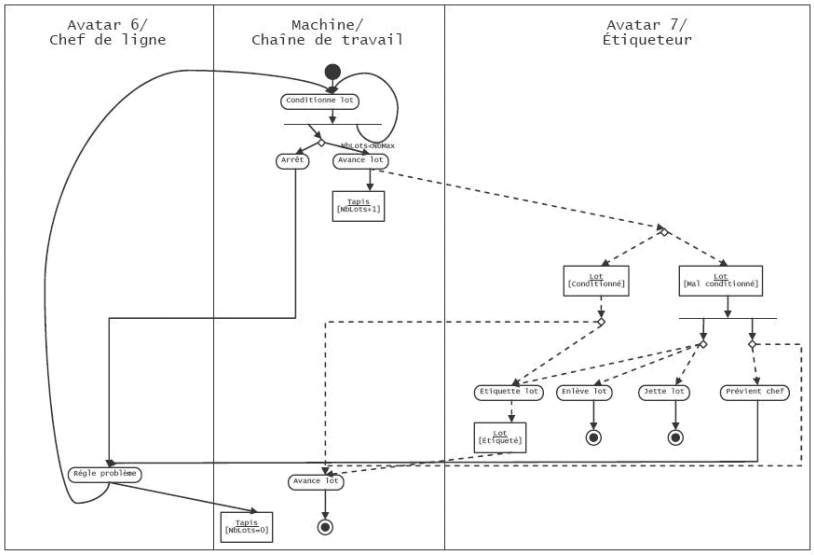


FIG. 6.7 – Diag. AUML d'activité des interactions entre agents (poste 7)

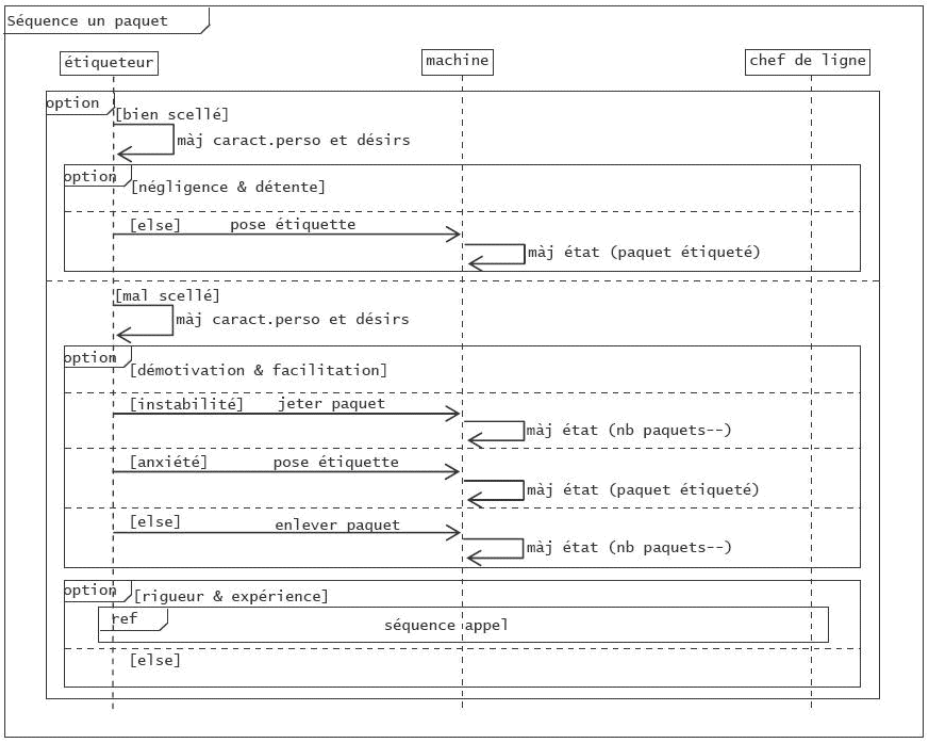


FIG. 6.8 – Diag. AUML de séquence des interactions entre agents (poste 7)

6.2.4 Implémentation dans Tc&Plus.Virtuel

L'ensemble de l'expertise recueillie auprès de psychologues a été formalisée dans les modèles AgentUML, puis implémentée dans la plateforme JADE. La plate-forme JADE [12] est un outil développé par CSELT en Java fournissant des bibliothèques de classes pour le développement d'agents et permettant d'animer ces agents au sein d'une plate-forme SMA. Cet outil est utilisé par une grande partie de la communauté multi-agents, et prend en charge certaines caractéristiques essentielles aux multi-agents telles que la communication et la concurrence sans pour autant contraindre exagérément la structure de l'agent.

Notre plateforme, appelée Tc&Plus.Virtuel (figure 6.9), a permis de réaliser un ensemble de simulations, afin de valider notre modèle auprès des psychologues d'APARTA. Les simulations permettent d'observer le fonctionnement de la chaîne de production et l'évolution de l'ambiance de travail en fonction de profils psychologiques des employés. L'outil offre donc un environnement d'aide à la décision aux psychologues pour la constitution d'équipes adaptées à leurs critères de réadaptation professionnelle.

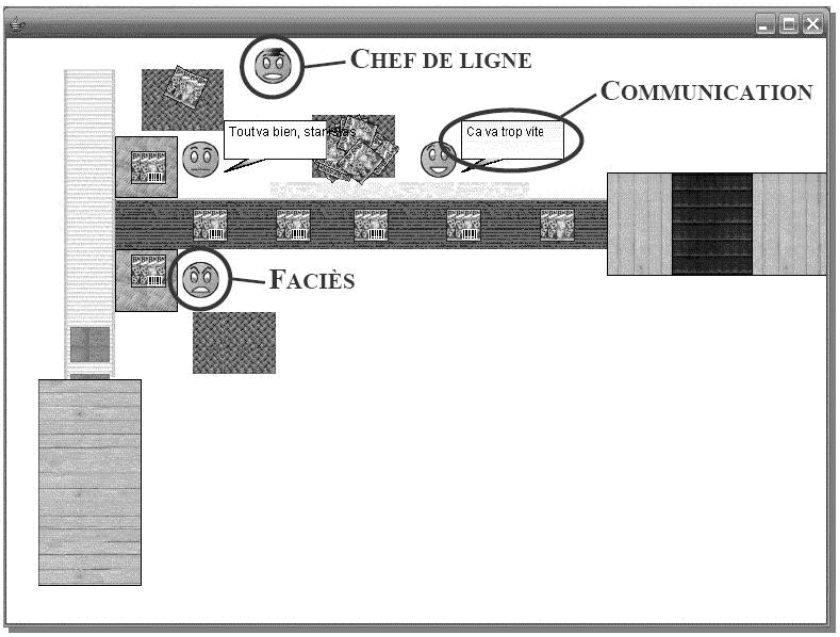


FIG. 6.9 – Exemple de simulation sur la plateforme Tc&Plus.Virtuel

6.3 Alignement d'ontologies

Dans cette section, nous présentons les travaux que nous avons réalisés sur l'alignement d'ontologies. Nous avons proposé une approche extensionnelle, asymétrique, et terminologique, originale, fondée sur une idée directrice : transposer le modèle des règles d'association, utilisé en fouille de données, afin de découvrir des alignements entre ontologies sous la forme tendances implicatives entre concepts. D'autre part, alors que les méthodes d'alignement disponibles sont orientées vers la production d'un alignement automatique, notre approche Aroma (Association Rule Ontology Matching) a été conçue dans une perspective d'aide à la décision, visant à offrir aux décideurs un environnement de découverte interactive d'alignements entre leurs ontologies.

6.3.1 Travaux connexes sur l'alignement

Les ontologies permettent de conceptualiser et partager des connaissances de manière structurée [93], et leur usage en gestion des connaissances tend à s'amplifier avec l'essor du Web sémantique. En pratique, les ontologies prennent des formes très variées depuis de simples taxonomies comme les systèmes de catégories (Yahoo, OpenDirectory), en passant par des systèmes de métadonnées interopérables (Dublin Core Metadata initiative) et allant jusqu'aux ontologies lourdes décrivant de véritables théories logiques (OWL, graphes conceptuels). On trouve donc souvent des ontologies différentes portant sur le même domaine qu'il serait souhaitable de fusionner ou de rapprocher. Dans cette optique, les techniques d'alignement visent à trouver des relations pertinentes entre deux ontologies (entre les classes/concepts, les relations, les propriétés...).

Dans la littérature, les travaux qui traitent de méthodes d'alignement reposent sur des techniques très différentes [136], comme l'apprentissage bayésien des probabilités jointes entre concepts [65], la classification conceptuelle [205], la fusion de schéma de bases de données [168], les modèles logiques en graphe conceptuels [78], la recherche de morphismes entre graphes représentant les ontologies [170]. La distinction entre ces travaux peut être faite au niveau des méthodes d'alignement utilisées. La classification proposée par Euzenat et Valtchev [69] distingue quatre familles de méthodes :

1. les *méthodes terminologiques* basées sur des mesures de similarité entre chaînes de caractères ou faisant intervenir une ressource terminologique externe ;
2. les *méthodes structurelles* comparant, d'une part, deux concepts à partir de mesures de similarité entre les constituants (attributs, propriétés)

des concepts ou à partir de leur position respective dans leur hiérarchie [174] ;

3. les *méthodes extensionnelles* comparant les concepts à partir de leur ensemble d'instances respectif [65] ;
4. les *méthodes sémantiques* basées sur un modèle sémantique théorique utilisé pour la comparaison des concepts [82, 78].

La plupart de ces travaux utilisent des relations symétriques de similarité. Pourtant, d'autres types de relations asymétriques peuvent être utilisées dans le but d'enrichir l'alignement produit. Par exemple, la recherche d'implications (généralisations) permet de trouver les concepts équivalents (exemple : si $auto \rightarrow voiture$ et $voiture \rightarrow auto$ alors $auto \leftrightarrow voiture$), mais elle permet aussi de découvrir si un concept est plus général (ou plus plus spécifique) qu'un autre. Parmi les méthodes prenant en compte la relation d'implication, nous pouvons citer S-MATCH [82]. Cette dernière évalue entre autre des relations d'équivalence et d'implication en s'appuyant sur un thésaurus (Wordnet).

6.3.2 Alignement avec Aroma

Aroma propose une méthode d'alignement extensionnelle, asymétrique et terminologique, basée sur la découverte de règles d'association entre deux ontologies. Notre méthode traite les relations de généralisation/subsomption et d'instanciation de concepts dans les ontologies. Elle a initialement été conçue pour l'alignement d'ontologies spécifiques : des taxonomies de documents constituées d'une hiérarchie de concepts, dont les concepts sont associés à des documents textuels partageant un vocabulaire commun. Puis, elle été étendue à l'alignement d'ontologies plus expressives formalisées en langage OWL.

Notre approche est divisée en deux phases consécutives qui utilisent toutes deux la mesure probabiliste d'écart à l'indépendance appelé intensité d'implication [83] :

1. *L'extraction terminologique*. Cette phase de prétraitement permet d'extraire dans les documents associés, un ensemble de termes pertinents pour chaque concept. Et ainsi d'instancier chaque concept par un ensemble de termes représentatifs.
2. *L'extraction d'implications*. Cette seconde phase extrait les règles d'association pertinentes entre les concepts sur la base des termes qui les représentent. La relation de généralisation/spécialisation de chaque hiérarchie est prise en compte afin de n'extraire les règles les plus générales et ainsi réduire la redondance.

6.3.2.1 Extraction terminologique

Notre objectif consiste à extraire des documents un ensemble de termes significatifs pour chacun des concepts d'une ontologie.

Mais, au préalable, nous devons extraire l'ensemble T des termes présents dans les documents associés à l'ontologie. Pour cela, nous avons utilisé les résultats de nos travaux précédents sur la validation d'expertise textuelle [56, 57, 58]. Nous proposons de réaliser une extraction des verbes et des termes binaires (termes composés de deux mots significatifs), à l'aide du logiciel ACABIT [50], sur des textes préalablement étiquetés (grammaticalement) et lemmatisés par la suite logicielle MontyLingua [165]. Les termes binaires présentent l'avantage d'être moins ambigus que des termes restreints à un seul mot.

Puis, l'idée principale de cette phase est la suivante : un terme $t \subseteq T$ sera significatif d'un concept c, si il existe peu de documents contenant le terme t qui ne sont pas associés au concept c ou à un de ses sous concepts. Pour cela, nous choisissons d'associer le terme t au concept c si la règle d'association $t \rightarrow c$ est significative, c'est-à-dire possède une valeur d'intensité d'implication supérieure à un seuil fixé ($\varphi(t \rightarrow c) > \varphi_t$).

La valeur d'intensité d'implication est définie par :

$$\varphi(t \rightarrow c) = 1 - Pr\left(N_{t\bar{c}} \leq n_{t\bar{c}}\right)$$

où $n_{t\bar{c}}$ représente le nombre observé de documents contenant le terme t qui ne sont pas associés au concept c, et $N_{t\wedge\bar{c}}$ représente un modèle aléatoire de $n_{t\bar{c}}$ sous hypothèse d'indépendance des descriptions t et c. Comme les phénomènes étudiés sont rares, nous modélisons la variable aléatoire $N_{t\bar{c}}$ par une loi de Poisson de paramètre $\lambda = n_t.n_{\bar{c}}/n$, où n_t est le nombre de documents contenant le terme t, $n_{\bar{c}}$ le nombre de documents non associés au concept c, et n le nombre total de documents.

A l'issue de cette première phase, nous disposerons de la liste $\gamma(c) \subseteq T$ des termes significatifs de chaque concept c d'une ontologie.

6.3.2.2 Extraction d'implications

Cette deuxième phase a pour objectif de découvrir des règles d'association binaires entre les concepts de deux ontologies distinctes. Une règle d'association $A \rightarrow B$ entre un concept A et B de deux ontologies distinctes sera retenue comme significative si les deux critères suivants sont vérifiés :

1. sa valeur d'intensité d'implication $\varphi(A \rightarrow B)$ est supérieur à un seuil φ_r fixé.

2. Il n'existe pas une règle d'association retenue qui permette de la déduire.

Le premier critère se traduit par $\varphi(A \rightarrow B) > \varphi_r$. L'intensité d'implication est définie par $\varphi(A \rightarrow B) = 1 - Pr\,(N_{A\overline{B}} \leq n_{A\overline{B}})$ sur l'ensemble T' des termes significatifs communs aux deux ontologies. $n_{A\overline{B}} = \mid \gamma(A) - \gamma(B) \mid$ représente le nombre observé de termes associés au concept A qui ne sont pas associés au concept B. $N_{A\overline{B}}$ est un modèle aléatoire sous hypothèse d'indépendance des descriptions A et B définit par une loi de Poisson de paramètre $\lambda = \mid \gamma(A) \mid . \mid T' - \gamma(B) \mid / \mid T' \mid$.

Le deuxième critère, inspiré des travaux sur les règles d'association généralisées [204], permet de réduire le nombre d'associations en éliminant les formes redondantes. Il se traduit par la formule $\forall X, \forall Y$ avec $A \leq X$ et $Y \leq B$, $\varphi(X \rightarrow Y) \leq \varphi(A \rightarrow B)$, signifiant que toute règle plus spécifique que $A \rightarrow B$ se voit attribuer une valeur d'intensité d'implication inférieure. Ce deuxième critère traduit la capacité d'une règle à générer d'autres règles.

6.3.2.3 Résultats expérimentaux

L'outil Aroma a été testé sur plusieurs jeux d'ontologies. La première série d'expérimentations a été effectuée sur deux jeux de test proposés dans [65]. Le premier jeu de test "Course Catalog" décrit les cours proposés par l'université de Cornell et l'université de Washington. Ce premier jeu de test a l'avantage d'être associé à un alignement manuel symétrique qui nous servira de référence. Les hiérarchies contiennent respectivement 166 et 176 concepts ainsi que 4360 et 6957 instances sous forme de documents textuels décrivant des cours. Les cours (instances) sont organisés en écoles et collèges, ensuite en départements et centres à l'intérieur de chaque collège. Le deuxième jeu de test "Company Profile" est issu des annuaires web Yahoo.com et Standard.com. Ces dernières hiérarchies contiennent respectivement 115 et 333 noeuds ainsi que 13634 et 9504 instances. On peut remarquer que la hiérarchie Standard.com a un découpage plus fin que celle de Yahoo.com. Les instances décrivent les activités d'entreprises. Les descriptions d'entreprises sont organisées en secteurs puis en industries.

Les résultats de cette étude sont détaillés dans [53]. On y montre que l'utilisation des règles d'association permet de découvrir des alignements portant du sens qui ne sont pas présents dans l'alignement manuel associé au premier jeu de test, et qui ne pourraient pas être découverts par les approches symétriques. On observe également une bonne précision des alignements découverts relativement aux alignements manuels, mais un rappel plus faible. Toutefois, il faut noter la difficulté de la comparaison de nos résultats aux alignements de références disponibles. En effet, la comparaison est biaisée

par la nature différentes des alignements comparés : asymétriques (implications) pour ceux que découvrons, et symétriques (équivalences) pour les alignements de référence.

L'outil Aroma a également été testé sur des ontologies OWL dans le cadre du banc d'essai de référence OAEI (Ontology Alignment Evaluation Initiative) data [1]. L'adaptation de notre approche, ainsi que la comparaison des résultats obtenus à ceux des approches GLUE et oPLMap, sont détaillés dans [53, 51, 54].

6.4 Synthèse des publications

L'ensemble des travaux réalisés dans le domaine de la gestion et de l'ingénierie des connaissances présentés dans ce chapitre ont été associés à des développements de logiciels. Ils ont été menés pour la plupart dans le cadre de collaborations avec des entreprises, et ont abouti à un transfert technologique. Ainsi, nous avons réalisé deux implémentations successives de l'approche serveur de connaissances : *SAMANTA*[2] avec La poste [182] [108], puis *Athanor* avec PerformanSE SA [74] [73] [103]. Une amélioration du placement des graphes sous-jacents aux représentations graphiques des modèles ontologiques a été proposée dans la thèse de Bruno Pinaud [185, 184]. Une modélisation d'agents émotionnels a été implémentée dans la plateforme multi-agent *TCplusVirtuel* [59] [60] [169]. Enfin, dans le cadre de la thèse de Jérôme David, notre approche d'alignement asymétrique d'ontologie a été implémentée dans l'outil *AROMA*[3], testée sur plusieurs jeux de références, et comparée à d'autres méthodes d'alignement [53] [54] [55] [52] [51].

En complément du caractère appliqué de ces travaux, nous avons également proposé un ensemble d'approches novatrices. Nous avons conçu (i) une approche *serveur de connaissances* incorporant des modèles ontologiques. Nous avons proposé (ii) une modélisation EST pour des *agents émotionnels* et leurs interactions, exprimée avec Agent UML. Nous avons conçu (iii) une méthode originale (terminologique, extensionnelle, asymétrique), afin de faciliter l'aide à la décision pour l'*alignement d'ontologie*.

[2]Système d'Aide à la MAiNtenance de Trieuses Automatiques
[3]Association Rule Ontology MAtching

Chapitre 7

Conclusion et Perspectives

Dans ce document, j'ai proposé un panorama de nos travaux de recherche sur le traitement des connaissances. Ces travaux ont principalement concerné deux domaines de recherche : l'extraction de connaissances dans les données, et la gestion et l'ingénierie des connaissances. J'ai également mis en relief deux préoccupations sous-jacentes : la première pour l'aide à la décision à travers la conception d'interfaces visuelles interactives centrées sur l'utilisateur, et la seconde pour l'application de nos travaux et la conception d'outils informatiques.

Nos principales contributions peuvent être ventilées en quatre volets : les travaux sur les mesures de qualité, l'approche fouille de règles, les ontologies pour la gestion des connaissances, et le développement d'outils.

Mesures de qualité. Nous avons publié un ensemble de travaux de *synthèse* sur les mesures de qualité, à travers (i) *deux ouvrages collectifs* [38][105], (ii) un recensement et une *classification* originale de 40 mesures [25, 23], et (iii) une série d'*études comparatives* sur plusieurs base de règles volumineuses, grâce à une approche originale de visualisation des corrélations entre mesures à l'aide *graphes* [120]. Nous avons (iv) développé *trois mesures de qualité* originales : deux mesures entropiques EII [30] et TIC [27], et une mesure statistique IPEE [25]. Nous avons également proposé (v) *plusieurs extensions* de la mesure statistique d'intensité d'implication selon les différents contextes de : données de séquences [17], variables ordinales [95] et floues [86], réduction de variables [48] et élimination des redondances [154], et enfin de cohésion de classes de règles [87].

Fouille de Règles. Nous avons proposé (i) une approche anthropocentrée originale pour la *fouille de règles*, basée sur un principe cognitif de fouille locale par *ciblage de règles*, et adossée à une *représentation graphique interactive* munie d'opérateurs de manipulation [152, 16]. Nous avons décliné cette approche en deux représentations interactives originales. D'une part, une re-

présentation par (ii) des *graphes de règles* dotée de capacités dynamiques [146] ; et d'autre part, une représentation par (ii) une *métaphore 3D* représentant des paysages de règles liés par une relation de voisinage [22].

Gestion de connaissances et ontologies. Nous avons conçu (i) une approche *serveur de connaissances* incorporant des modèles ontologiques [73] [103]. Nous avons proposé (ii) une modélisation EST pour des *agents émotionnels* et leurs interactions, exprimée en Agent UML [59]. Nous avons conçu (iii) une méthode originale (terminologique, extensionnelle, asymétrique), afin de faciliter l'aide à la décision pour l'*alignement d'ontologies* [54, 51].

Outils. Les travaux sur les mesures de qualité nous ont amené à implémenter (i) deux outils téléchargeables sur le web : *ARVAL*[1] pour le calcul des mesures, et la plateforme *ARQAT*[2] pour l'étude expérimentale du comportement des mesures. (ii) Deux implémentations de l'approche de fouille de règles ont été réalisées et testées sur des données : *Felix* pour les graphes[155], et *ARVIS*[3] pour les métaphores 3D [22]. Nous avons réalisé (iii) deux implémentations successives de l'approche serveur de connaissances : *SAMANTA*[4] avec La poste [182] [108], et *Athanor* avec PerformanSE SA [74]. (iv) Les agents émotionnels ont été implémentés dans la plateforme multi-agent *TC-plusVirtuel* [59], [60], [169]. Enfin, (v) l'alignement asymétrique d'ontologie a été implémenté dans l'outil *AROMA*[5] [53], [54], [51].

7.1 Perspectives

J'envisage de prolonger de cette activité en la projetant dans la direction des deux verrous scientifiques majeurs souvent mentionnés en ECD :

- le passage à l'échelle sur les *masses de connaissances*,
- et la *fouille de connaissances* vues comme des données complexes.

De mon point de vue, cette perspective a aussi l'avantage d'allier des enjeux *académiques* et des enjeux *d'applications en entreprises*. Je discerne une clé, sous-jacente à ces 2 verrous, cachée dans l'établissement de *ponts scientifiques* entre la *gestion et ingénierie des connaissances* et la *fouille de données*.

1. **Passage à l'échelle sur les masses de connaissances.** Sur les données très volumineuses, ou "masses de données", la plupart des approches classiques voient leurs performances s'effondrer dramatiquement. Ceci provient du fait que la complexité des algorithmes sous-

[1] Association Rule VAlidation
[2] Association Rule QuAlity Tool
[3] Association Rule VISualisation
[4] Système d'Aide à la MAiNtenance de Trieuses Automatiques
[5] Association Rule Ontology MAtching

jacents dépend du nombre d'enregistrements (i.e. plus d'un million) ; et dépend, souvent plus fortement encore, du nombre de variables traitées (i.e. plus d'un milliers). Deux types de solutions, une sur les données, et l'autre sur les algorithmes, sont envisageables. La première, bien traitée dans la littérature, consiste à réduire la taille des données par échantillonnage des enregistrements et/ou par réduction des variables, mais elle incorpore un risque de perte d'information. La seconde, moins traitée, consiste à réduire la complexité des algorithmes soit en recherchant la linéarité, soit en distribuant les calculs sur une infrastructure virtuelle (grid computing ou grille de calcul), soit enfin en quittant les approches globales pour passer à des approches locales incorporant des connaissances du domaine, ou des heuristiques d'exploration.

Cette seconde piste me semble la plus prometteuse. Nous avons commencé à l'aborder dans notre approche anthropocentrée de fouille de règles, où les connaissances de l'utilisateur dirigent implicitement, telle une heuristique, un algorithme de fouille local. J'envisage de prolonger cette approche, en créant un "pont scientifique" avec l'ingénierie des connaissances, qui permettra d'utiliser explicitement les connaissances de l'utilisateur après les avoir extraites et formalisée. Les travaux sur les mesures de qualité subjectives sont de cet ordre, mais ils se heurtent à la difficulté de l'extraction préalable des connaissances du domaine.

En aval du processus d'ECD, le passage à l'échelle se traduit aussi sur les "masses de connaissances" lorsque les algorithmes produisent de grandes quantités d'informations (comme les règles d'association). Comme précédemment, les approches locales restent intéressantes. Mais, dans ce contexte, j'envisage une autre perspective qui s'appuierait sur un couplage avec l'ingénierie des connaissances et plus précisément avec les grilles de connaissances. En effet, les grilles de connaissances (ou semantic grid [42]) constituent une approche récente, issue du croisement des grilles de calcul et des techniques du web sémantique, qui a l'avantage d'offrir une infrastructure collaborative de stockage et de services interopérables ouvrant l'accès au traitement de très grandes masses de connaissances. Ainsi, tirant bénéfice du support offert par ces grilles, je propose de développer un service d'annotation sémantique des connaissances, où l'utilisateur pourrait restituer une sémantique personnalisée aux connaissances intéressantes une fois celles-ci repérées dans la masse des résultats.

2. **Fouille de connaissances.** Stimulés par l'apparition et sur le web et dans les SBDB du langage XML, nouveau standard interopérable pour les données semi-structurées, les données complexes se banalisent. La

nature complexe des données est liée à leur structure (graphes, dimensions spatiale et temporelles, ...), à leur hétérogénéité (textes, données, multimédia), et enfin à leur localisation (multi-sources).

Ayant toujours à l'esprit le pont évoqué entre l'ECD et la gestion des connaissances, je discerne trois perspectives autour de la fouille des données complexes. (1) Partant du constat que les connaissances peuvent êtres considérées comme des données à structure complexe, la première perspective consiste à développer des algorithmes spécifiques de fouille de données dans des bases de connaissances (composées d'ontologies OWL, de règles, ...), afin de découvrir des relations (méta-connaissances ?) au seins des connaissances stockées. (2) En transposant les techniques de fouille pour l'analyse des traces d'utilisation dans les log des sites web, on peut s'intéresser au développement d'algorithmes de fouille spécifiques pour analyse les traces d'usage des connaissances dans les bases de connaissances. Cette perspective constitue une extension naturelle de mes travaux sur le serveur de connaissances Athanor. (3) Elle peut aussi être appliquées à l'analyse des traces d'activité laissées par les agents émotionnels décrits précédemment.

Bibliographie

[1] Aggarwal, C. C. Towards effective and interpretable data mining by visual interaction. *SIGKDD Explorations 3*, 2 (2002), 11–22.

[2] Aggarwal, C. C., and Yu, P. S. Mining associations with the collective strength approach. *IEEE Transactions on Knowledge and Data Engineering 13*, 6 (2001), 863–873.

[3] Agrawal, R., Arning, A., Bollinger, T., Mehta, M., Shafer, J., and Srikant, R. The quest data mining system. In *Proceedings of the second ACM SIGKDD international conference on knowledge discovery and data mining* (1996), AAAI Press, pp. 244–249.

[4] Agrawal, R., Imielienski, T., and Swami, A. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on management of data* (1993), ACM Press, pp. 207–216.

[5] Agrawal, R., and Srikant, R. Fast algorithms for mining association rules. In *VLDB'94, Proceedings of 20th International Conference on Very Large Data Bases* (1994), J. B. Bocca, M. Jarke, and C. Zaniolo, Eds., Morgan Kaufmann, pp. 487–499.

[6] Alavi, M., and Leidner, D. Review : Knowledge management and knowledge management systems : Conceptual foundations and research issues. *Mis Quarterly 25*, 1 (2001), 107–136.

[7] Andrews, K. Visualising cyberspace : information visualisation in the Harmony internet browser. In *Proceedings of the 1995 IEEE symposium on Information Visualization* (1995), IEEE Computer Society, pp. 97–104.

[8] Azé, J. Une nouvelle mesure de qualité pour l'extraction de pépites de connaissances. *Revue des Sciences et Technologies de l'Information 17*, 1-3 (2003), 171–182. Actes des journées Extraction et Gestion des Connaissances (EGC) 2003.

[9] BANDHARI, I. Attribute focusing : machine-assisted knowledge discovery applied to software production process control. *Knowledge Acquisition journal 6*, 3 (1994), 271–294.

[10] BARTHÉLEMY, J.-P., AND MULLET, E. A model of selection by aspects. *Acta Psychologica 79*, 1 (1992), 1–19.

[11] BAYARDO, R.-J., AND AGRAWAL, R. Mining the most interesting rules. In *5^{th} International Conference on Knowledge Discovery and Data Mining* (1999), ACM SIGKDD, ACM Press, pp. 145–154.

[12] BELLIFEMINE, F., POGGI, A., AND RIMASSA, G. Jade-a fipa-compliant agent framework. In *PAAM 99* (1999), pp. 97–108.

[13] BERTI-EQUILLE, L., AND GUILLET, F., Eds. *Actes du 1er atelier Qualité des Données et des Connaissances (QDC), Conférence Extraction et Gestion des Connaissances (EGC'05)* (Paris, Janvier 2005), Association EGC.

[14] BERTI-EQUILLE, L., AND GUILLET, F., Eds. *Actes du 2ème atelier Qualité des Données et des Connaissances (QDC), Conférence Extraction et Gestion des Connaissances (EGC'06)* (Lille, Janvier 2006), Association EGC.

[15] BERTIN, J. *Sémiologie graphique*. Gauthier-Villars, 1967. (3e édition en 1999 aux Editions de l'Ecole des Hautes Etudes en Sciences Sociales).

[16] BLANCHARD, J. *Un système de visualisation pour l'extraction, l'évaluation, et l'exploration interactive de règles d'association*. PhD thesis, Universite de Nantes, 2005.

[17] BLANCHARD, J., GUILLET, F., AND BRIAND, H. L'intensité d'implication entropique pour la recherche de règles de prédiction intéressantes dans les séquences de pannes d'ascenseurs. *Extraction des Connaissances et Apprentissage (ECA), Hermès Science Publication 1(4)* (2002), 77–88.

[18] BLANCHARD, J., GUILLET, F., AND BRIAND, H. Exploratory visualization for association rule rummaging. In *Proceedings of the fourth international workshop on Multimedia Data Mining in conjunction with ACM KDD'2003* (2003), pp. 107–114.

[19] BLANCHARD, J., GUILLET, F., AND BRIAND, H. A user-driven and quality oriented visualization for mining association rules. In *Proc. of the 3rd IEEE Int. Conf. on Data Mining, ICDM'2003* (Melbourne, Florida, USA, 2003), IEEE Computer Society Press, pp. 493–497.

[20] BLANCHARD, J., GUILLET, F., AND BRIAND, H. A virtual reality environment for knowledge mining. In *Proc. of the Human Center Process*

conf. on Distributed Decision Making and Man-Machine Cooperation (HCP'03) (Luxembourg, 2003), pp. 175–179.

[21] Blanchard, J., Guillet, F., and Briand, H. Une visualisation orientée qualité pour la fouille anthropocentrée de règles d'association. *In Cognito - Cahiers Romans de Sciences Cognitives 1*, 3 (2004), 79–100. ISSN 1267-8015.

[22] Blanchard, J., Guillet, F., and Briand, H. Interactive visual exploration of association rules with the rule focusing methodology. *Knowledge and Information Systems* (2006), To appear. to appear.

[23] Blanchard, J., Guillet, F., Briand, H., and Gras, R. Assessing rule interestingness with a probabilistic measure of deviation from equilibrium. In *11th international symposium on Applied Stochastic Models and Data Analysis (ASMDA'2005)* (2005), ENST, pp. 191–200.

[24] Blanchard, J., Guillet, F., Briand, H., and Gras, R. IPEE : Indice Probabiliste d'Ecart à l'Equilibre pour l'évaluation de la qualité des règles. In *Actes de l'atelier Qualité des Données et des Connaissances en conjonction avec la conférence EGC'05* (2005), pp. 26–34.

[25] Blanchard, J., Guillet, F., Briand, H., and Gras, R. Une version discriminante de l'Indice Probabiliste d'Ecart à l'Equilibre pour mesurer la qualité des règles. *Revue Quaderni di Ricerca in Didattica 15*, 2 (2005), 131–138.

[26] Blanchard, J., Guillet, F., Gras, R., and Briand, H. Mesurer la qualité des règles et de leurs contraposées avec le taux informationnel tic. *Revue Nationale des Technologies de l'Information (RNTI), E2 1* (2004), 287–298. ISBN 2-85428-633-2.

[27] Blanchard, J., Guillet, F., Gras, R., and Briand, H. Using information-theoretic measures to assess association rule interestingness. In *ICDM'05, Proceedings of the 5th IEEE International Conference on Data Mining* (2005), IEEE Computer Society Press, pp. 66–73.

[28] Blanchard, J., Guillet, F., Rantière, F., and Briand, H. Vers une représentation graphique en réalité virtuelle pour la fouille interactive de règles d'association. *Numéro spécial revue Extraction et Gestion des Connaissances 17*, 1,2,3 (2003), 105–118.

[29] Blanchard, J., Kuntz, P., Guillet, F., and Gras, R. Améliorer la mesure de l'étonnement statistique des règles à l'aide de la contraposée. Rapport de l'as gafodonnées/gafoqualité, Rapport de l'AS GafoDonnées/GafoQualité - STIC - CNRS, 2002.

[30] Blanchard, J., Kuntz, P., Guillet, F., and Gras, R. *Implication intensity : from the basic statistical definition to the entropic version.* Chapman and Hall/CRC Press, 2003, pp. 473–485. chapter 28.

[31] Blanchard, J., Kuntz, P., Guillet, F., and Gras, R. Mesure de qualité des règles d'association par l'intensité d'implication entropique. *Mesures de qualité pour la fouille de données, Numéro spécial Revue des Nouvelles Technologies de l'Information (RNTI) E1*, 1 (2004), 33–44. ISBN 2-85428-646-4.

[32] Blanchard, J., Lehn, R., Guillet, F., and Kuntz, P. Des graphes à la réalité virtuelle pour l'extraction adaptative de règles d'association. In *Actes de l'atelier Visualisation et Extraction Adaptative des Connaissances, conférence EGC 2003* (Lyon, 2003).

[33] Blanchard, J., Poulet, F., Guillet, F., and Kuntz, P. Highly interactive data mining with virtual reality. In *Proc. 5th Virtual Reality International Conf. (VRIC'2003)* (Laval, 2003), ISTIA Innovation, pp. 221–228. ISBN 2-9515730-2-2.

[34] Brachman, J., and Anand, T. The process of knowledge discovery in databases : a human-centered approach. In *Advances in knowledge discovery and data mining*, U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Eds. AAAI/MIT Press, 1996, pp. 37–58.

[35] Braga, D., Campi, A., Klemettinen, M., and Lanzi, P. L. Mining association rules from XML data. In *Proceedings of the fourth international conference on data warehousing and knowlege discovery (DaWaK 2002)* (2002), vol. 2454 of *Lecture Notes in Computer Science*, Springer-Verlag, pp. 21–30.

[36] Bratman, M. *Intentions, Plans and Practical Reason.* Harvard University Press, 1987.

[37] Briand, H., and Guillet, F., Eds. *Extraction et Gestion des Connaissances, numéro spécial Revue Extraction des Connaissances et Apprentissage (ECA)*, vol. vol. 1, n° 1-2 of *Revue Extraction des Connaissances et Apprentissage (ECA)*. Hermès Science Publications, 2001. ISBN 2-7462-0216-6.

[38] Briand, H., Sebag, M., Gras, R., and Guillet, F., Eds. *Mesures de qualité pour la fouille de données, Numéro spécial Revue des Nouvelles Technologies de l'Information (RNTI)*, vol. E-1 of *Numéro spécial Revue des Nouvelles Technologies de l'Information (RNTI)*. Cépaduès Edition, 2004. ISBN 2-85428-646-4.

[39] Brin, S., Motwani, R., and Silverstein, C. Beyond market baskets : Generalizing association rules to correlations. In *ACM SIGMOD Conference on Management of Data* (1997), ACM Press, pp. 265–276.

[40] Brin, S., Motwani, R., Ullman, J. D., and Tsur, S. Dynamic itemset counting and implication rules for market basket data. In *ACM SIGMOD International Conference on Management of Data* (1997), ACM Press, pp. 255–264.

[41] Brunk, C., Quelly, J., and Kohavi, R. Mineset : An integrated system for data mining. In *Proceedings of the third ACM SIGKDD international conference on knowledge discovery and data mining* (1997), D. Heckerman, H. Mannila, D. Pregibon, and R. Uthurusamy, Eds., AAAI Press, pp. 135–138.

[42] Cannataro, M., and Talia, D. Semantics and knowledge grids : Building the next-generation grid. *IEEE Intelligent Systems Vol. 19* (2004), 56–63.

[43] Clark, P., and Niblett, T. The CN2 induction algorithm. *Machine Learning 3*, 4 (1989), 261–283.

[44] Cohen, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20 (1960), 37–46.

[45] Corby, O., Dieng-Kuntz, R., and Faron-Zucker, C. Querying the semantic web with the corese search engine. In *16th European Conference on Artificial Intelligence (ECAI'2004), subconference PAIS'2004* (Valencia, 2004, 22-27 August 2004), R. L. de Mantaras and L. S. eds, Eds., IOS Press, pp. 705–709.

[46] Corby, O., and Faron, C. Corese : A corporate semantic web engine. In *Proceedings of the International Workshop on Real World RDF and Semantic Web Applications, 11th International World Wide Web Conference* (Hawai, USA, May 2002).

[47] Couturier, R., and Gras, R. CHIC : traitement de données avec l'analyse implicative. *Revue des Nouvelles Technologies de l'Information E-3* (2005), 679–684. Actes des journées Extraction et Gestion des Connaissances (EGC) 2005.

[48] Couturier, R., Gras, R., and Guillet, F. Reducing the number of variables using implicative analysis. In *Conference of the International Federation of Classification Societies (IFCS'04)* (Chicago, USA, July 15-18 2004), Springer Verlag, pp. 277–285.

[49] Daconta, M. C., Obrst, L. J., and Smith., K. T. *The Semantic Web : A Guide to the Future of XML, Web Services, and Knowledge Management.* Kluwer, 2003.

[50] DAILLE, B. Conceptual structuring through term variations. In *ACL 2003 Workshop on Multiword Expressions : Analysis, Acquisition and Treatment* (2003), F. Bond, A. Korhonen, D. MacCarthy, and A. Villacicencio, Eds., pp. 9–16.

[51] DAVID, J., GUILLET, F., AND BRIAND, H. Mapping directories and owl ontologies with aroma. In *ACM Conference on Information and Knowledge Management (CIKM'06)* (november 2006), ACM. To appear.

[52] DAVID, J., GUILLET, F., GRAS, R., AND BRIAND, H. Alignement de taxonomies documentaires : une méthode asymétrique et extensionnelle. In *conférence Ingénierie des Connaissances (IC2006)* (Nantes, Juin 2006), p. A paraître. Poster.

[53] DAVID, J., GUILLET, F., GRAS, R., AND BRIAND, H. Alignement extensionnel et asymétrique de hiérarchies conceptuelles par découverte d'implications entre concepts. *Revue des Nouvelles Technologies de l'Information E*, 6 (2006), 151–162.

[54] DAVID, J., GUILLET, F., GRAS, R., AND BRIAND, H. Conceptual hierarchies matching : an approach based on discovery of implication rules between concepts. In *17th European Conference on Artificial Intelligence (ECAI)* (2006), IOS Press, pp. 357–361. ISBN 1-58603-642-4.

[55] DAVID, J., GUILLET, F., GRAS, R., AND BRIAND, H. An interactive, asymmetric and extensional method for matching conceptual hierarchies. In *Open INTEROP Workshop On "Enterprise Modelling and Ontologies for Interoperability" (INTEROP-EMOI06)* (Luxembourg, June 5-6 2006), p. To appear.

[56] DAVID, J., GUILLET, F., PHILIPPE, V., BRIAND, H., AND GRAS, R. Validation d'une expertise textuelle par une méthode de classification basée sur l'intensité d'implication. In *Actes de l'atelier Qualité des Données et des Connaissances de la conférence Extraction et Gestion des Connaissances (EGC'05)* (Paris, 18 janvier 2005), pp. 50–57.

[57] DAVID, J., GUILLET, F., PHILIPPE, V., BRIAND, H., AND GRAS, R. Validation d'une expertise textuelle par une méthode de classification basée sur l'intensité d'implication. *Revue Quaderni di Ricerca in Didattica 15*, 2 (2005), 157–162. ISSN 1592-5137.

[58] DAVID, J., GUILLET, F., PHILIPPE, V., AND GRAS, R. Implicative statistical analysis applied to clustering of terms taken from a psychological text corpus. In *11th international symposium on Applied Stochastic Models and Data Analysis (ASMDA'2005)* (Brest, France,

May 17-20 2005), J.Janssen and P. (eds), Eds., pp. 201–208. ISBN 2-908849-15-1.

[59] DAVIET, S., DESMIER, H., BRIAND, H., GUILLET, F., AND PHILIPPÉ, V. A system of emotional agents for decision support. In *IEEE/WIC/ACM international conference on Intelligent Agent Technology (IAT'05)* (2005), A. Skowron, J.-P. Barthes, L. Jain, R. Sun, P. Morizet-Mahoudeaux, J. Liu, and N. Zhong, Eds., IEEE Computer Society Press, pp. 711–717.

[60] DESMIER, H., GUILLET, F., FLOREA, A. M., BRIAND, H., AND PHILIPPE, V. Modélisation d'un agent émotionnel en uml et rdf. *Revue des Nouvelles Technologies de l'Information (RNTI-E-3) E : Extraction et Gestion des Connaissances*, 3 (2005), 637–642. ISBN 2-85428-677-4.

[61] DI-BATTISTA, G., EADES, P., TAMASSIA, R., AND TOLLIS, I.-G. *Graph drawing – Algorithms for the visualization of graphs*. Prentice-Hall, 1999.

[62] DICE, L. Measures of the amount of ecologic association between species. *Ecology*, 26 (1945), 297–302.

[63] DIENG, R., CORBY, O., GIBOIN, A., AND RIBIERE, M. Methods and tools for corporate knowledge management. *International Journal of Human-Computer Studies, special issue on Knowledge Management 51* (1999), 567–598.

[64] DIENG-KUNTZ, R., CORBY, O., GANDON, F., GIBOIN, A., GOLEBIOWSKA, J., MATTA, N., AND RIBIÈRE, M. *Méthodes et outils pour la gestion des connaissances*. Dunod, 2001. 2^{eme} édition.

[65] DOAN, A., MADHAVAN, J., DOMINGOS, P., AND HALEVY, A. *Ontology Matching : a machine learning approach*. Springer-Velag, 2004, pp. 397–416.

[66] EARL, M. Knowledge management strategies : toward a taxonomy. *J. of Management Information Systems 18*, 1 (2001), 215–233.

[67] EL-NASR, M. S., IOERGER, T. R., AND YEN, J. A web of emotions. In *workshop on Emotion-Based Agent Architecture (EBAA 99) at the* 3^{rd} *International Conference On Autonomous Agents (Agents 99)* (2000).

[68] ERMINE, J.-L. *Les systèmes de connaissances*. Hermès, Paris, deuxième édition, 2000.

[69] EUZENAT, J., AND VALTCHEV, P. An integrative proximity measure for ontology alignment. In *Semantic Integration Workshop, Second International Semantic Web Conference (ISWC-03)* (2003).

[70] FAYYAD, U., GRINSTEIN, G., AND WIERSE, A., Eds. *Information visualization in data mining and knowledge discovery*. Morgan Kaufmann, 2001.

[71] FLEURY, L. *Découverte de connaissances dans une base de données de gestion des ressources humaines*. PhD thesis, Université de Nantes, 1996.

[72] FLOREA, A. M., AND KALISZ, E. Behavior anticipation based on beliefs, desires and emotions. In *6^{th} international conference on Computing Anticipatory Systems (CASYS 03)* (2003).

[73] FOLLUT, D., GALLETTI, F., AND GUILLET, F. Les arbres de décision interactifs (adi). In *14ème Forum National de la Maintenance* (Paris, Novembre 2002), éditions AFIM.

[74] FOLLUT, D., GUILLET, F., PHILIPPÉ, J., AND VANDEKERCKHOVE, P. Athanor - un système pour la capitalisation et le déploiement de connaissances de diagnostic. *Extraction des Connaissances et Apprentissage (ECA) 1*, 1,2,3 (2001), 315–324. ISBN 2-7462-0216-6.

[75] FRAWLEY, W., PIATETSKY-SHAPIRO, G., AND MATHEUS, C. Knowledge discovery in databases : an overview. *AI Magazine 14*, 3 (1992), 57–70.

[76] FREITAS, A. A. On objective measures of rule surprisingness. In *Proceedings of the second European conference on principles of data mining and knowledge discovery (PKDD'98)* (1998), J. Zytkow and M. Quafafou, Eds., vol. 1510 of *Lecture Notes in Artificial Intelligence*, Springer-Verlag, pp. 1–9.

[77] FREITAS, A. A. On rule interestingness measures. *Knowledge-Based Systems Journal 12(5-6)* (1999), 309–315.

[78] FÜRST, F., AND TRICHET, F. Aligner des ontologies lourdes : une méthode basée sur les axiomes. In *16èmes journées francophones d'ingénierie des connaisssances* (2005), pp. 121–132.

[79] FULE, P., AND RODDICK, J. F. Experiences in building a tool for navigating association rule result sets. In *CRPIT'04 : Proceedings of the second Australasian workshop on information security, data mining, web intelligence, and software internationalisation* (2004), J. Hogan, P. Montague, M. Purvis, and C. Steketee, Eds., Australian Computer Society, Inc., pp. 103–108.

[80] GANASCIA, J.-G. Charade : apprentissage de bases de connaissances. In *Induction symbolique et numérique à partir de données*, Y. Kodratoff and E. Diday, Eds. Cépaduès Editions, 1991, pp. 309–326.

[81] Gavrilov, M., Anguelov, D., Indyk, P., and Motwani, R. Mining the stock market : which measure is best ? *KDD'00, Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining* (2000), 487–496.

[82] Giunchiglia, F., Shvaiko, P., and Yatskevich, M. S-match : an algorithm and an implementation of semantic matching. In *European Semantic Web Symposium* (2004), LNCS 3053, pp. 61–75.

[83] Gras, R., Almouloud, S., Bailleuil, M., Lahrer, A., Polo, M., Ratsimba-Rajohn, H., and Totohasina, A. *L'implication statistique : nouvelle méthode exploratoire de données*. Edition de la pensée sauvage, 1996.

[84] Gras, R., Couturier, R., Bernadet, M., Blanchard, J., Briand, H., Kuntz, P., Guillet, F., Lehn, R., and Peter, P. Un exemple de mesure de qualité : l'implication statistique. Rapport de l'as gafodonnées/gafoqualité, Rapport de l'AS GafoDonnées/GafoQualité - STIC - CNRS, 2002.

[85] Gras, R., Couturier, R., Blanchard, J., Briand, H., Kuntz, P., and Peter, P. Quelques critères pour une mesure de qualité de règles d'association. *numéro spécial Mesures de qualité pour la fouille de données, Revue des Nouvelles Technologies de l'Information E-1* (2004), 3–31. numéro spécial Mesures de qualité pour la fouille de données.

[86] Gras, R., Couturier, R., Guillet, F., and Spagnolo, F. Extraction de règles en incertain par la méthode implicative. In *Actes de l'atelier Qualité des Données et des Connaissances (QDC), conférence Extraction et Gestion des Connaissances (EGC'05)* (Paris, 18 janvier 2005), pp. 19–25.

[87] Gras, R., David, J., Régnier, J.-C., and Guillet, F. Typicalité et contribution des sujets et des variables supplémentaires en Analyse Statistique Implicative. *Revue des Nouvelles Technologies de l'Information E-6* (2006), 359–370. ISSN 1764-1667.

[88] Gras, R., Guillet, F., Peter, P., and Philippé, J. Apprentissage automatique et implication : mise en oeuvre sur un espace d'apprentissage en ressources humaines. In *4èmes journées de la Société Francophone de Classification* (Vannes, 19-20 Septembre 1996).

[89] Gras, R., Guillet, F., and Philippe, J. Réduction des colonnes d'un tableau de données par quasi-équivalence entre variables. *Extraction des Connaissances et Apprentissage 1(4)* (2002), 197–202.

[90] GRAS, R., KUNTZ, P., AND BRIAND, H. Les fondements de l'analyse statistique implicative et quelques prolongements pour la fouille de données. *Mathématiques et Sciences Humaines 39*, 154-155 (2001), 9–29.

[91] GRAS, R., KUNTZ, P., COUTURIER, R., AND GUILLET, F. Une version entropique de l'intensité d'implication pour les corpus volumineux. *Extraction des Connaissances et Apprentissage 1*, 1-2 (2001), 69–80.

[92] GRAS, R., PETER, P., BAQUEDANO, S., AND PHILIPPÉ, J. Structuration de comportements de réponse à un questionnaire par des méthodes multi-dimensionnelles. *Extraction des Connaissances et Apprentissage (ECA), Hermès 17(1-3)* (2003), 105–118.

[93] GRUBER, T. R. A translation approach to portable ontology specifications. *Knowledge Acquisition 5*, 2 (1993), 199–220.

[94] GUILLAUME, S. *Traitement des données volumineuses, mesures et algorithmes d'extraction de règles d'association et règles ordinales*. Thèse de doctorat, Université de Nantes, 2000.

[95] GUILLAUME, S., AND GUILLET, F. Une généralisation des règles d'association par l'intensité d'implication ordinale. In *7èmes Journées de la Société Francophone de Classification (SFC)* (Nancy, Septembre 1999), pp. 77–86.

[96] GUILLAUME, S., GUILLET, F., AND PHILIPPE, J. Contribution of the integration of intensity of implication into the algoritm proposed by agrawal. In *14th European Meeting on Cybernetics and Systems Research (EMCSR'98)* (Vienna, Austria, April 14-17 1998), vol. 2, Austrian Society of Cybernetic Studies, pp. 805–810. ISBN 3-85206-139-3.

[97] GUILLAUME, S., GUILLET, F., AND PHILIPPE, J. Improving the discovery of association rules with intensity of implication. In *Second European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD'98)* (1998), vol. LNCS 1510, Springer, pp. 318–327.

[98] GUILLAUME, S., GUILLET, F., AND PHILIPPE, J. L'intensité d'implication pour la découverte de règles d'association "pertinentes". In *1ères journées Ré-ingéniérie des Systèmes d'Information (RSI'98)* (Lyon, 1-2 Avril 1998), vol. LNCS 1510, pp. 36–43.

[99] GUILLET, F. *Contributions à la maîtrise de qualité d'un processus industriel par apprentissage des stratégies de contrôle de l'opérateur*. PhD thesis, ENST-Br, Université de Rennes I, Décembre 1995.

[100] GUILLET, F., Ed. *1ère journée Qualité des Connaissances (GafoQualité)* (Nantes, Mars 2002), Polytech'Nantes.

[101] GUILLET, F. Rapport d'activité et projets de recherche sur la qualité des connaissances. Rapport de l'as gafodonnées/gafoqualité, Rapport de l'AS GafoDonnées/GafoQualité - STIC - CNRS, 2002.

[102] GUILLET, F. Mesures de la qualité des connaissances en ECD. In *Actes des tutoriels, 4ème Conférence Extraction et Gestion des Connaissances (EGC'04)* (2004), pp. 1–60.

[103] GUILLET, F., FOLLUT, D., AND PHILIPPÉ, J. Athanor : Une approche pour la gestion des connaissances de maintenance sur les systèmes complexes. In *Première journée Sstèmes d'information pour l'aide à la décision en ingénierie système, JESIADIS 2002* (Brest, 2002), ENSIETA, pp. 41–54.

[104] GUILLET, F., GRAS, R., KUNTZ, P., AND BRIAND, H. Mesure de qualité de règles d'association par analyse implicative. In *1ere journée Qualité des Connaissances (GafoQualite)* (Nantes, France, 2002), p. 5.

[105] GUILLET, F., AND HAMILTON, H. *Quality measures in Data Mining*. Studies in Computational Intelligence. Springer, 2006. ISBN 3540449116.

[106] GUILLET, F., KUNTZ, P., AND LEHN, R. A genetic algorithm for visualizing networks of association rules. In *Proceedings of the 12^{th} International Conference on Industrial and Engineering Applications of Artificial Intelligence, IEA/AIE'99* (1999), vol. 1611 of *Lecture Notes in Computer Science*, Springer Verlag, pp. 145–154.

[107] GUILLET, F., AND LENCA, P. Invited session : Quality in data mining. 11th International Symposium on Applied Stochastic Models and Data Analysis (ASMDA'2005), May 17-20 2005.

[108] GUILLET, F., AND VANDEKERCKHOVE, P. Samanta : a knowledge server. In *Workshop on Knowledge Management - Theory and Application in conjunction with PKDD'2000* (Lyon, September 12 2000), J.-L. Ermine, Ed., pp. 19–28.

[109] HAN, J. Towards on-line analytical mining in large databases. *SIGMOD Record 27*, 1 (1998), 97–107. DBMiner.

[110] HAN, J., AN, A., AND CERCONE, N. Cviz : an interactive visualization system for rule induction. In *AI'00 : Proceedings of the thirteenth Biennial Conference of the Canadian Society on Computational Studies of Intelligence* (2000), Springer-Verlag, pp. 214–226.

[111] HAN, J., FU, Y., WANG, W., KOPERSKI, K., AND ZAIANE, O. DMQL : a data mining query language for relational databases. In *Proceedings of the 1996 SIGMOD workshop on research issues on data mining and knowledge discovery (DMKD)* (1996).

[112] Han, J., Hu, X., and Cercone, N. A visualization model of interactive knowledge discovery systems and its implementations. *Information Visualization 2*, 2 (2003), 105–125.

[113] Hao, M. C., Dayal, U., Hsu, M., Sprenger, T., and Gross, M. H. Visualization of directed associations in e-commerce transaction data. In *Proceedings of VisSym 2001*, pp. 185–192.

[114] Hilderman, R. J., and Hamilton, H. J. *Knowledge Discovery and Measures of Interestingness*. Kluwer Academic Publishers, 2001.

[115] Hofmann, H., and Wilhelm, A. Visual comparison of association rules. *Computational Statistics 16*, 3 (2001), 399–415.

[116] Holland, J. *Adaptation in natural and artificial systems*, 2ème ed. M.I.T. Press, 1992.

[117] Holland, J., Holyoak, K., Nisbett, R., and Thagard, P. *Induction : Processes of inference, learning and discovery*. MIT Press, 1986.

[118] Hussain, F., Liu, H., Suzuki, E., and Lu, H. Exception rule mining with a relative interestingness measure. In *Knowledge Discovery and Data Mining, Current Issues and New Applications, 4th Pacific-Asia Conference (PAKDD'00)* (2000), pp. 86–97.

[119] Huynh, X.-H. *Interestingness measures for association rules in a KDD process : postprocessing of rules with ARQAT tool*. PhD thesis, Ecole polytechnique de l'université de Nantes, 2006.

[120] Huynh, X.-H., Guillet, F., Blanchard, J., Kuntz, P., Gras, R., and Briand, H. *A graph-based clustering approach to evaluate interestingness measures : a tool and a comparative study*, vol. Quality measures in Data Mining of *Studies in Computational Intelligence*. Springer-Verlag, 2006, ch. 2, p. To appear. ISBN 3540449116.

[121] Huynh, X.-H., Guillet, F., and Briand, H. Arqat : une plateforme d'analyse exploratoire pour la qualité des règles d'association. In *Actes de l'atelier Qualité des Données et des Connaissances de la conférence Extraction et Gestion des Connaissances (EGC'05)* (Paris, 18 janvier 2005), pp. 58–68.

[122] Huynh, X.-H., Guillet, F., and Briand, H. Arqat : une plateforme exploratoire pour la qualité des règles d'association : apports pour l'analyse implicative. In *Actes de la 3ème conférence sur l'Analyse Statistique Implicative (ASI'05)* (Palerme, Octobre 2005), pp. 6–8. A paraitre.

[123] Huynh, X.-H., Guillet, F., and Briand, H. ARQAT : an exploratory analysis tool for interestingness measures. In *11th International Symposium on Applied Stochastic Models and Data Analysis (ASMDA'05)* (2005), pp. 334–344.

[124] Huynh, X.-H., Guillet, F., and Briand, H. Clustering interestingness measures with positive correlation. In *ICEIS'05, Proceedings of 7th International Conference on Enterprise Information Systems* (2005), vol. 2, pp. 248–253.

[125] Huynh, X.-H., Guillet, F., and Briand, H. A data analysis approach for evaluating the behavior of interestingness measures. In *DS'05, Proceedings of the 8th International Conference on Discovery Science* (2005), vol. LNAI 3735, Springer, pp. 330–337.

[126] Huynh, X.-H., Guillet, F., and Briand, H. Comparaison des mesures d'intérêt de règles d'association : une approche basée sur des graphes de corrélation. *Revue des Nouvelles Technologies de l'Information (RNTI-E-6) E : Extraction et Gestion des Connaissances*, 6 (2006), 549–560. ISSN 1764-1667.

[127] Huynh, X.-H., Guillet, F., and Briand, H. Discovering the stable clusters between interestingness measures. In *8th International Conference on Enterprise Information Systems (ICEIS'06)* (Cyprus, 2006), pp. 196–201. ISBN 972-8865-41-4.

[128] Huynh, X.-H., Guillet, F., and Briand, H. Evaluating interestingness measures with correlation graph. In *19th international conference on Industrial, Engineering & Other Applications of Applied Intelligent Systems (IEA-AIE'06)* (Annecy (France), 2006), lncs 4031, Springer, pp. 312–321. ISBN 3-540-35453-0.

[129] Huynh, X.-H., Guillet, F., and Briand, H. Extracting representative measures for the post-processing of association rules. In *4th international IEEE conference on Computer Sciences : Research & Innovation - Vision for the Future (RIVF'06)* (Ho-chiminh Ville (Vietnam), 2006), IEEE Computer Society Press, pp. 99–106.

[130] Huynh, X.-H., Guillet, F., and Briand, H. Extraction de mesures d'intérêt représentatives pour le post-traitement des règles d'association. In *Atelier Qualité des Données et des Connaissances de la conférence Extraction et Gestion des Connaissances (DKQ-EGC'05)* (Lille (France), 2006), pp. 45–54.

[131] Huynh, X.-H., Guillet, F., and Briand, H. A graph-based approach for comparing interestingness measures. In *First IEEE International Conference on Engineering of Intelligent Systems (IEEE*

ICEIS'06) (Islamabad, Pakistan, 2006), IEEE press, p. To appear. A paraître.

[132] IMIELINSKI, T., AND MANNILA, H. A database perspective on knowledge discovery. *Communications of the ACM 39*, 11 (1996), 58–64.

[133] IMIELINSKI, T., AND VIRMANI, A. MSQL : a query language for database mining. *Data Mining and Knowledge Discovery 3*, 4 (1999), 373–408.

[134] JACCARD, P. Etude comparative de la distribution florale dans une portion des Alpes et du Jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37 (1901), 547–579.

[135] JAROSZEWICZ, S., AND SIMOVICI, D. A. A general measure of rule interestingness. In *Proceedings of the fifth European conference on principles of data mining and knowledge discovery (PKDD'01)* (2001), vol. LNCS 2168, Springer-Verlag, pp. 253–265.

[136] KALFOGLOU, Y., AND SCHORLEMMER, M. Ontology mapping : the state of the art. *Knowedge Engineering Review 18*, 1 (2003), 1–31.

[137] KAPOOR, A., MOTA, S., AND PICARD, R. Towards a learning companion that recognizes affect. In *In Proc. of Emotional and Intelligent : The Tangled knot of social cognition* (2001), AAAI Fall Symposium.

[138] KEIM, D. A. Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics 8*, 1 (2002), 1–8.

[139] KLEMETTINEN, M., MANNILA, H., RONKAINEN, P., TOIVONEN, H., AND VERKAMO, A. I. Finding interesting rules from large sets of discovered association rules. In *Proceedings of the third international conference on information and knowledge management (CIKM 1994)* (1994), ACM Press, pp. 401–407.

[140] KLEMETTINEN, M., MANNILA, H., AND TOIVONEN, H. Interactive exploration of discovered knowledge : a methodology for interaction and usability studies. Tech. rep., University of Helsinki, 1996. TR C-1996-3.

[141] KODRATOFF, Y. Extraction de connaissances à partir des données et des textes. In *Actes des journées sur la fouille dans les données par la méthode d'analyse statistique implicative* (2000), Presses de l'Université de Rennes 1, pp. 151–165.

[142] KONONENCO, I. On biases in estimating multi-valued attributes. *IJCAI'95, Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence* (1995), 1034–1040.

[143] KOPANAKIS, I., AND THEODOULIDIS, B. Visual data mining and modeling techniques. In *Proceedings of the KDD-2001 workshop on visual data mining* (2001).

[144] KULCZYNSKI, S. Die pflanzenassoziationen der pieninen. *Bulletin International de l'Académie Polonaise des Sciences et des Lettres. Classe des Sciences Mathématiques et Naturelles*, suppl. II (1927), 57–203. série B.

[145] KUNTZ, P., GUILLET, F., LEHN, R., AND BRIAND, H. A user-driven process for mining association rules. In *Proceedings of the fourth European conference on principles of data mining and knowledge discovery (PKDD-2000)* (2000), Springer-Verlag, pp. 483–489.

[146] KUNTZ, P., GUILLET, F., LEHN, R., AND BRIAND, H. Vers un processus d'extraction de règles d'association centré sur l'utilisateur. *In Cognito, Revue francophone internationale en sciences cognitives 1*, 20 (2001), 13–26. ISSN 1267-8015.

[147] KUNTZ, P., LEHN, R., AND BRIAND, H. Dynamic rule graph drawing by genetic search. In *Proceedings of the IEEE International Conference on System Man and Cybernetics* (2000).

[148] KUNTZ, P., LEHN, R., AND BRIAND, H. Interactive rule-network layout with a genetic algorithm in a knowledge discovery process. In *Proceedings of Data Mining* (2000), vol. 2, Wessex Institute of Technology, WIT Press, pp. 435–444.

[149] KUNTZ, P., LEHN, R., GUILLET, F., AND PINAUD, B. *Visualisation en Extraction de Connaissances, numéro spécial Revue Nationale des Technologies de l'Information (RNTI)*, vol. E1. Cépaduès, 2006, ch. Découverte interactive de règles d'association via une interface visuelle, pp. 113–126.

[150] LALLICH, S., AND TEYTAUD, O. Evaluation et validation de l'intérêt des règles d'association. *numéro spécial Mesures de Qualité pour la Fouille de Données, Revue des Nouvelles Technologies de l'Information (RNTI) E-1* (2004), 193–217.

[151] LAVRAC, N., FLACH, P. A., AND ZUPAN, B. Rule evaluation measures : a unifying view. In *ninth International Workshop on Inductive Logic Programming (ILP'99)* (1999), LNAI 1634, Springer-Verlag, pp. 174–185.

[152] LEHN, R. *Un système interactif de visualisation et de fouille de règles pour l'extraction de connaissances dans les bases de données*. PhD thesis, Université de Nantes, 2000.

[153] LEHN, R., GUILLET, F., AND BRIAND, H. Eliminating redundant knowledge in an association rule-based system : an algorithm. In *14th European Meeting on Cybernetics and Systems Research (EMCSR'98)* (Vienna, Austria, April 14-17 1998), vol. 2, Austrian Society of Cybernetic Studies, pp. 793–798. ISBN 3-85206-139-3.

[154] LEHN, R., GUILLET, F., AND BRIAND, H. Qualité d'un ensemble de règles : élimination des règles redondantes. *"Mesures de Qualité pour la Fouille de Données", Numéro spécial Revue Nationale des Technologies de l'Information (RNTI) E1* (2004), 141–168. ISBN 2-85428-646-4.

[155] LEHN, R., GUILLET, F., KUNTZ, P., BRIAND, H., AND PHILIPPÉ, J. Felix : An interactive rule mining interface in a kdd process. In *10^{th} Mini-Euro Conference, Human Centered Processes (HCP'99)* (September 1999), P. Lenca, Ed., ENST-Br, pp. 169–174.

[156] LEHN, R., GUILLET, F., KUNTZ, P., BRIAND, H., AND PHILIPPÉ, J. Félix : un outil interactif d'aide à la fouille de connaissances s'appuyant sur l'intensité d'implication. In *Fouille dans les données par la méthode d'analyse statistique implicative* (September 2000), R. Gras and M. Bailleul, Eds., Association pour la recherche en Didactique des Mathématiques, pp. 51–58. ISBN 2-9516505-0-7.

[157] LEHN, R., KUNTZ, P., AND GUILLET, F. Sur un problème de tracé de graphes posé par un processus interactif d'extraction de règles d'associations. In *3ème congrès de la société Francaise de Recherche Opérationnelle et d'Aide à la Décision (ROADEF'2000)* (Nantes, 26-28 Janvier 2000), pp. 151–156.

[158] LENCA, P., MEYER, P., VAILLANT, B., PICOUET, P., AND LALLICH, S. Evaluation et analyse multicritère des mesures de qualité des règles d'association. *numéro spécial Mesures de Qualité pour la Fouille de Données, Revue des Nouvelles Technologies de l'Information (RNTI) E-1* (2004), 219–246. numéro spécial Mesures de qualité pour la fouille de données.

[159] LERMAN, I. C. *Classification et analyse ordinale des données*. Dunod, 1981.

[160] LIU, B., AND HSU, W. Post-analysis of learned rules. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI-96)* (1996), AAAI Press, pp. 828–834.

[161] LIU, B., HSU, W., CHEN, S., AND MA, Y. Analyzing the subjective interestingness of association rules. *IEEE Intelligent Systems 15*, 5 (2000), 47–55.

[162] Liu, B., Hsu, W., and Ma, Y. Pruning and summarizing the discovered associations. In *KDD'99 : Proceedings of the fifth ACM SIGKDD international conference on knowledge discovery and data mining* (1999), ACM Press, pp. 125–134.

[163] Liu, B., Hsu, W., Mun, L.-F., and Lee, H.-Y. Finding interesting patterns using user expectations. *IEEE Transactions on Knowledge and Data Engineering 11(6)* (1999), 817–832.

[164] Liu, B., Hsu, W., Wang, K., and Chen, S. Visually aided exploration of interesting association rules. In *PAKDD'99 : Proceedings of the Third Pacific-Asia Conference on Methodologies for Knowledge Discovery and Data Mining* (1999), Springer-Verlag, pp. 380–389.

[165] Liu, H. Montylingua : An end-to-end natural language processor with common sense, 2004.

[166] Loevinger, J. A systematic approach to the construction and evaluation of tests of ability. *Psychological Monographs 61*, 4 (1947).

[167] Ma, Y., Liu, B., and Wong, C. K. Web for data mining : organizing and interpreting the discovered rules using the web. *SIGKDD Explorations 2*, 1 (2000), 16–23.

[168] Madhavan, J., Bernstein, P. A., and Rahm, E. Generic schema matching with cupid. In *The VLDB Journal* (2001), pp. 49–58.

[169] Marinica, C., Guillet, F., and Briand, H. Représentation d'interactions entre agents par des règles ris formalisées en rdf. In *conférence Ingénierie des Connaissances (IC2006)* (Nantes, Juin 2006), p. A paraître. Poster.

[170] Melnik, S., Garcia-Molina, H., and Rahm, E. Similarity flooding : A versatile graph matching algorithm and its application to schema matching. In *the 18th International Conference on Data Engineering (ICDE'02)* (2002), IEEE Computer Society, pp. 117–128.

[171] Meo, R., Psaila, G., and Ceri, S. An extension to SQL for mining association rules. *Data Mining and Knowledge Discovery 2*, 2 (1998), 195–224.

[172] Montgomery, H. Decision rules and the search for a dominance structure : towards a process model of decision making. In *Analysing and aiding decision processes*, P. Humphreys, O. Svenson, and A. Vari, Eds. Elsevier Science Publishers, 1983, pp. 343–369.

[173] Mosteller, F. Association and estimation in contingency tables. *Journal of the American Statistical Association 63*, 321 (1968), 1–28.

[174] Noy, N. F., and Musen, M. A. Prompt : Algorithm and tool for automated ontology merging and alignment. In *the 17th National Conference on Artificial Intelligence and 12th Conference on Innovative Applications of Artificial Intelligence* (2000), AAAI Press / The MIT Press, pp. 450–455.

[175] Ochiai, A. Zoogeographic studies on the soleoid fishes found in japan and its neighbouring regions. *Bulletin of the Japanese Society of Scientific Fisheries*, 22 (1957), 526–530.

[176] Odell, J., Parunak, H. V. D., and Bauer, B. Extending uml for agents. In *Agent-oriented information systems workshop at the* 17^{th} *National Conference on Artificial Intelligence (NCAI 00)* (2000), pp. pp. 3–17.

[177] Ortony, A., Clore, G., and Collins, A. *The cognitive structure of emotions*. Cambridge University Press, 1998.

[178] Ouni, A., and Dudezert, A. Etat de l'art des approches de définition du sytème de gestion des connaissances. In 9^{e} *colloque de l'AIM : Systèmes d'information : perspectives critiques* (2004).

[179] Padmanabhan, B., and Tuzhilin, A. A belief-driven method for discovering unexpected patterns. In *KDD'98, Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining* (1998), pp. 94–100.

[180] Pearson, K. Mathematical contributions to the theory of evolution : regression, heredity and panmixia. *Philosophical Transactions of the Royal Society Of London series A*, 187 (1896), 253–318.

[181] Pham, D., Dimov, S., and Peat, B. Intelligent product manuals. *Journal of Engineering Manufacture, Proc. of the Institution of Mechanical Engineers 214*, B5 (2000), 411–419. Partie B.

[182] Philippé, J., Guillet, F., Follut, D., and Vandekerckhove, P. Un serveur de connaissances dans un contexte de maintenance appliquée aux machine de tri postal. In *Journées Internationales Ingénierie des systèmes et NTIC, NimesTIC'2000* (Nîmes, France, 11-13 Spetembre 2000), pp. 30–35.

[183] Piatetsky-Shapiro, G. Discovery, analysis, and presentation of strong rules. In *Knowledge Discovery in Databases*, G. Piatetsky-Shapiro and W. J. Frawley, Eds. AAAI/MIT Press, 1991, pp. 229–248.

[184] Pinaud, B., Kuntz, P., Guillet, F., and Philippé, V. Graphical visualization in the knowledge management system atanor. In *International Conference on Knowledge Management (I-Know'06)* (2006),

Journal of Universal Computer Science (J.UCS), Springer, pp. 481–488. ISSN 0948-6968.

[185] Pinaud, B., Kuntz, P., Guillet, F., and Philippé, V. Visualisation en gestion des connaissances, développement d'un nouveau modèle graphique Graph'Atanor. *Revue des Nouvelles Technologies de l'Information E*, 6 (2006), 311–322. ISSN 1764-1667.

[186] Prax, J. *Le Manuel du Knowledge Management*. Dunod, 2003.

[187] Purchase, H. Which aesthetic has the greatest effect on human understanding ? In *Proceedings of Graph Drawing'97* (1997), vol. 1353 of *Lecture Notes in Computer Sciences*, Springer Verlag, pp. 248–261.

[188] Rainsford, C. P., and Roddick, J. F. Visualisation of temporal interval association rules. In *Proceedings of the second international conference on intelligent data engineering and automated learning (IDEAL 2000)* (2000), Springer-Verlag, pp. 91–96.

[189] Rao, A., and Georgeff, M. Modeling rational agents within a bdi-architecture. In 2^{nd} *International Conference on Principles of Knowledge Representation and Reasoning (KR 91)* (1991), Morgan Kaufmann Publishers Inc, pp. 473–484.

[190] Rickel, J., Marsella, S., Gratch, J., Hil, R., Traum, D., and Swartout, W. Steve goes to bosnia : towards a new generation of virtual humans for interactive experiences. *IEEE Intelligent Systems 17(4)* (2002), 32–38.

[191] Ritschard, G., Zighed, D. A., and Nicoloyannis, N. Maximisation de l'association par regroupement de lignes ou de colonnes d'un tableau croisé. *Mathématiques et Sciences Humaines 39*, 154-155 (2001), 81–97.

[192] Rogers, D., and Tanimoto, T. A computer program for classifying plants. *Science*, 132 (1960), 1115–1118.

[193] Ross, S. M. *Introduction to probability and statistics for engineers and scientists*. Wiley, 1987.

[194] Russel, P., and Rao, T. On habitat and association of species of anopheline larvae in south-eastern madras. *Journal of the Malaria Institute of India*, 3 (1940), 153–178.

[195] Saporta, G. *Probabilités, analyse des données, et statistique*. Editions Technip, 1990.

[196] Schreiber, G., Wielinga, B., de Hoog, R., Akkermans, H., and de Velde, W. V. CommonKADS : a comprehensive methodology for KBS development. *IEEE Expert 9*, 6 (1994), 28–37.

[197] SEBAG, M., AND SCHOENAUER, M. Generation of rules with certainty and confidence factors from incomplete and incoherent learning bases. In *Proceedings of the European knowledge acquisition workshop EKAW'88* (1988), Gesellschaft für Mathematik und Datenverarbeitung mbH, pp. 28.1–28.20.

[198] SHANNON, C., AND WEAVER, W. *The mathematical theory of communication.* University of Illinois Press, 1949.

[199] SHNEIDERMAN, B. Inventing discovery tools : combining information visualization with data mining. *Information Visualization 1*, 1 (2002), 5–12.

[200] SILBERSCHATZ, A., AND TUZHILIN, A. What makes patterns interesting in knowledge discovery systems. *IEEE Transactions on Knowledge and Data Engineering 8*, 6 (1996), 970–974.

[201] SIMON, H. *Models of Thought.* Yale University Press, 1979.

[202] SMYTH, P., AND GOODMAN, R. M. An information theoretic approach to rule induction from databases. *IEEE Transactions on Knowledge and Data Engineering 4*, 4 (1992), 301–316.

[203] SOKAL, R., AND MICHENER, C. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 38 (1958), 1409–1438.

[204] SRIKANT, R., AND AGRAWAL, R. Mining generalized association rules. In *21st International Conference on Very Large Databases (VLDB'95)* (1995), pp. 407–419.

[205] STUMME, G., AND MAEDCHE, A. FCA-MERGE : Bottom-up merging of ontologies. In *IJCAI* (2001), pp. 225–234.

[206] SUZUKI, E., AND ZYTKOW, J. M. Unified algorithm for undirected discovery of exception rules. *International Journal of Intelligent Systems 20*, 7 (2005), 673–691.

[207] TAN, P.-N., AND KUMAR, V. Interestingness measures for association patterns : a perspective. In *Proceedings of the KDD-2000 workshop on postprocessing in machine learning and data mining* (2000).

[208] TAN, P.-N., KUMAR, V., AND SRIVASTAVA, J. Selecting the right interestingness measure for association patterns. *8th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'02)* (2002), 32–41.

[209] TAN, P.-N., KUMAR, V., AND SRIVASTAVA, J. Selecting the right objective measure for association analysis. *Information Systems 29*, 4 (2004), 293–313.

[210] THEIL, H. On the estimation of relationships involving qualitative variables. *American Journal of Sociology 76* (1970), 103–154.

[211] TUZHILIN, A., AND ADOMAVICIUS, G. Handling very large numbers of association rules in the analysis of microarray data. In *KDD'02 : Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining* (2002), ACM Press, pp. 396–404.

[212] VAILLANT, B., LENCA, P., AND LALLICH, S. Etude expérimentale de mesures de qualité de règles d'association. *Revue des Nouvelles Technologies de l'Information (RNTI) E-2* (2004), 341–352.

[213] WILKINSON, L. *The Grammar of Graphics*. 2005.

[214] WONG, P. C., WHITNEY, P., AND THOMAS, J. Visualizing association rules for text mining. In *Proceedings of the 1999 IEEE symposium on information visualization* (1999), IEEE Computer Society, pp. 120–123.

[215] WOOLDRIDGE, M., AND JENNINGS, N. R. *Intelligent agent : theory and practice*, vol. vol. 10, issue 2. Knowledge Engineering Review, Cambridge University Press, 1995.

[216] YULE, G. On the association of attributes in statistics. *Philosophical Transactions of the Royal Society of London series A*, 194 (1900), 257–319.

[217] ZHAO, Y., AND KARYPIS, G. Criterion functions for document clustering : experiments and analysis. Tech. rep., Deparment of Computer Science, University of Minnesota, 2001. Technical Report TR01-40.

ANNEXES

Annexe A

A graph-based clustering approach to evaluate interestingness measures : a tool and a comparative study [120]

Chapter 2 in book : F. Guillet and H. Hamilton (eds), *Quality measures in Data Mining, Studies in Computational Intelligence*, Springer, 2007.

A graph-based clustering approach to evaluate interestingness measures : a tool and a comparative study

Xuan-Hiep Huynh, Fabrice Guillet, Julien Blanchard, Pascale Kuntz, Henri Briand, and Régis Gras

LINA CNRS 2729 - Polytechnic School of Nantes University, La Chantrerie BP 50609 44306 Nantes cedex 3, France {1stname.name}@polytech.univ-nantes.fr

Summary. Finding interestingness measures to evaluate association rules has become an important knowledge quality issue in KDD. Many interestingness measures may be found in the literature, and many authors have discussed and compared interestingness properties in order to improve the choice of the most suitable measures for a given application. As interestingness depends both on the data structure and on the decision-maker's goals, some measures may be relevant in some context, but not in others. Therefore, it is necessary to design new contextual approaches in order to help the decision-maker select the most suitable interestingness measures. In this paper, we present a new approach implemented by a new tool, ARQAT, for making comparisons. The approach is based on the analysis of a correlation graph presenting the clustering of objective interestingness measures and reflecting the post-processing of association rules. This graph-based clustering approach is used to compare and discuss the behavior of thirty-six interestingness measures on two prototypical and opposite datasets: a highly correlated one and a lowly correlated one. We focus on the discovery of the stable clusters obtained from the data analyzed between these thirty-six measures.

1 Introduction

As the number of discovered rules increases, end-users, such as data analysts and decision makers, are frequently confronted with a major challenge: how to validate and select the most interesting of those rules. Over the last decade the Knowledge Discovery in Databases (KDD) community has recognized this challenge – often referred to as interestingness – as an important and difficult component of the KDD process (Klemettinen et al. [15], Tan et al. [30]). To tackle this problem, the most commonly used approach is based on the construction of Interestingness Measures (IM).

In defining association rules, Agrawal et al. [1] [2] [3], introduced two IMs : support and confidence. These are well adapted to Apriori algorithm con-

straints, but are not sufficient to capture the whole aspects of the rule interestingness. To push back this limit, many complementary IMs have been then proposed in the literature (see [5] [14] [30] for a survey). They can be classified in two categories [10]: subjective and objective. Subjective measures explicitly depend on the user's goals and his/her knowledge or beliefs. They are combined with specific supervised algorithms in order to compare the extracted rules with the user's expectations [29] [24] [21]. Consequently, subjective measures allow the capture of rule novelty and unexpectedness in relation to the user's knowledge or beliefs. Objective measures are numerical indexes that only rely on the data distribution.

In this paper, we present a new approach and a dedicated tool ARQAT (Association Rule Quality Analysis Tool) to study the specific behavior of a set of 36 IMs in the context of a specific dataset and in an exploratory analysis perspective, reflecting the post-processing of association rules. More precisely, ARQAT is a toolbox designed to help a data-analyst to capture the most suitable IMs and consequently, the most interesting rules within a specific ruleset.

We focus our study on the objective IMs studied in surveys [5] [14] [30]. The list of IMs is added with four complementary IMs (Appendix A): Implication Intensity (II), Entropic Implication Intensity (EII), TIC (information ratio modulated by contra-positive), and IPEE (probabilistic index of deviation from equilibrium). Furthermore, we present a new approach based on the analysis of a correlation graph (CG) for clustering objective IMs.

This approach is applied to compare and discuss the behavior of 36 IMs on two prototypical and opposite datasets: a strongly correlated one (mushroom dataset [23]) and a lowly correlated one (synthetic dataset). Our objective is to discover the stable clusters and to better understand the differences between IMs.

The paper is structured as follows. In Section 2, we present related works on objective IMs for association rules. Section 3 presents a taxonomy of the IMs based on two criteria: the "subject" (deviation from independence or equilibrium) of the IMs and the "nature" of the IMs (descriptive or statistical). In Section 4, we introduce the new tool ARQAT for evaluating the behavior of IMs. In Section 5, we detail the correlation graph clustering approach. And, Section 6 is dedicated to a specific study on two prototypical and opposite datasets in order to extract the stable behaviors.

2 Related works on objective IMs

The surveys on the objective IMs mainly address two related research issues : *(1)* defining a set of principles or properties that lead to the design of a good IM, *(2)* comparing the IM behavior from a data-analysis point of view. The results yielded can be useful to help the user select the suitable ones.

Considering the principles of a good IM issue, Piatetsky-Shapiro [25] introduced the Rule-Interest, and proposed three underlying principles for a good IM on a rule $a \rightarrow b$ between two itemsets a and b : 0 value when a and b are independent, monotonically increasing with a *and* b, monotonically decreasing with a or b. Hilderman and Hamilton [14] proposed five principles: minimum value, maximum value, skewness, permutation invariance, transfer. Tan et al. [30] defined five interestingness principles: symmetry under variable permutation, row/column scaling invariance, anti-symmetry under row/column permutation, inversion invariance, null invariance. Freitas [10] proposed an "attribute surprisingness" principle. Bayardo and Agrawal [5] concluded that the most interesting rules according to some selected IMs must reside along a support/confidence border. The work allows for improved insight into the data and supports more user-interaction in the optimized rule-mining process. Kononenko [19] analyzed the biases of eleven IMs for estimating the quality of multi-valued attributes. The values of information gain, J-measure, Gini-index, and relevance tend to linearly increase with the number of values of an attribute. Zhao and Karypis [33] used seven different criterion functions with clustering algorithms to maximize or minimize a particular one. Gavrilov et al. [11] studied the similarity measures for the clustering of similar stocks. Gras et al. [12] discussed a set of ten criteria: increase, decrease with respect to certain expected semantics, constraints for semantics reasons, decrease with trivial observations, flexible and general analysis, discriminative residence with the increment of data volume, quasi-inclusion, analytical properties that must be countable, two characteristics of formulation and algorithms.

Some of these surveys also address the related issue of the IM comparison by adopting a data-analysis point of view. Hilderman and Hamilton [14] used the five proposed principles to rank summaries generated from databases and used sixteen diversity measures to show that: six measures matched five proposed principles, and nine remaining measures matched at least one proposed principle. By studying twenty-one IMs, Tan et al. [30] showed that an IM cannot be adapted to all cases and use both a support-based pruning and standardization methods to select the best IMs; they found that, in some cases many IMs are highly correlated with each other. Eventually, the decision-maker will select the most suitable measure by matching the five proposed properties. Vaillant et al. [31] evaluated twenty IMs to choose a user-adapted IM with eight properties: asymmetric processing of a and b for an association rule $a \rightarrow b$, decrease with n_b, independence, logical rule, linearity with $n_{a\bar{b}}$ around 0^+, sensitivity to n, easiness to fix a threshold, intelligibility. Finally, Huynh et al. [16] introduced the first result of a new clustering approach for classifying thirty-four IMs with a correlation analysis.

3 A taxonomy of objective IMs

In this section, we propose a taxonomy of the objective IMs (details in Appendixes A and B) according to two criteria: the "subject" (deviation from independence or equilibrium), and the "nature" (descriptive or statistical). The conjunction of these criteria seems to us essential to grasp the meaning of the IMs, and therefore to help the user choose the ones he/she wants to apply.

In the following, we consider a finite set T of transactions. We denote an association rule by $a \rightarrow b$ where a and b are two disjoint itemsets. The itemset a (respectively b) is associated with a transaction subset $A = T(a) = \{t \in T, a \subseteq t\}$ (respectively $B = T(b)$). The itemset $\overline{a}$ (respectively $\overline{b}$) is associated with $\overline{A} = T(\overline{a}) = T - T(a) = \{t \in T, a \not\subseteq t\}$ (respectively $\overline{B} = T(\overline{b})$). In order to accept or reject the general trend to have b when a is present, it is quite common to consider the number $n_{a\overline{b}}$ of negative examples (contra-examples, counter-examples) of the rule $a \rightarrow b$. However, to quantify the "surprisingness" of this rule, consider some definitions are functions of $n = |T|$, $n_a = |A|$, $n_b = |B|$, $n_{\overline{a}} = |\overline{A}|$, $n_{\overline{b}} = |\overline{B}|$.

Let us denote that, for clarity, we also keep the probabilistic notations $p(a)$ (respectively $p(b)$, $p(a \textit{ and } b)$, $p(a \textit{ and } \overline{b})$) as the probability of a (respectively b, a *and* b, a *and* $\overline{b}$). This probability is estimated by the frequency of a: $p(a) = \frac{n_a}{n}$ (respectively $p(b) = \frac{n_b}{n}$, $p(a \textit{ and } b) = \frac{n_{ab}}{n}$, $p(a \textit{ and } \overline{b}) = \frac{n_{a\overline{b}}}{n}$).

3.1 Subject of an IM

Generally speaking, an association rule is more interesting when it is supported by lots of examples and few negative examples. Thus, given n_a, n_b and n, the interestingness of $a \rightarrow b$ is minimal when $n_{a\overline{b}} = \min(n_a, n_{\overline{b}})$ and maximal when $n_{a\overline{b}} = \max(0, n_a - n_b)$. Between these extreme situations, there exist two significant configurations in which the rules appear non-directed relations and therefore can be considered as neutral or non-existing: the independence and the equilibrium. In these configurations, the rules are to be discarded.

Independence

Two itemsets a and b are independent if $p(a \textit{ and } b) = p(a) \times p(b)$, i.e. $n.n_{a\overline{b}} = n_a n_{\overline{b}}$. In the independence case, each itemset gives no information about the other, since knowing the value taken by one of the itemsets does not alter the probability distribution of the other itemset: $p(b \backslash a) = p(b \backslash \overline{a}) = p(b)$ and $p(\overline{b} \backslash a) = p(\overline{b} \backslash \overline{a}) = p(\overline{b})$ (the same for the probabilities of a and $\overline{a}$ given b or $\overline{b}$). In other words, knowing the value taken by an itemset lets our uncertainty about the other itemset intact. There are two ways of deviating from the independent situation: either the itemsets a and b are positively correlated ($p(a \textit{ and } b) > p(a) \times p(b)$), or they are negatively correlated ($p(a \textit{ and } b) < p(a) \times p(b)$).

Equilibrium

We define the equilibrium of a rule $a \rightarrow b$ as the situation where examples and negative examples are equal in numbers: $n_{ab} = n_{a\overline{b}} = \frac{1}{2}n_a$ [7]. In this situation, the itemset a is as concomitant with b as with $\overline{b}$ in the data. So a rule $a \rightarrow b$ at equilibrium is as directed towards b as towards $\overline{b}$. There are two ways of deviating from this equilibrium situation: either a is more concomitant with b than with $\overline{b}$, or a is more concomitant with $\overline{b}$ than with b.

Deviation from independence and from equilibrium

As there exist two different notions of neutrality, the objective interestingness of association rules must be measured from (at least) two complementary points of view: the deviation from independence, and the deviation from equilibrium. These are what we call the two possible subjects for the rule IMs. These deviations are directed in favor of examples and in disfavor of negative examples.

Definition 1. An IM m evaluates a deviation from independence if the IM has a fixed value at the independence:

$$m(n, n_a, n_b, \frac{n_a n_{\overline{b}}}{n}) = constant$$

Definition 2. An IM m evaluates a deviation from equilibrium if the IM has a fixed value at the equilibrium:

$$m(n, n_a, n_b, \frac{n_a}{2}) = constant$$

Independence is a function of four parameters n, n_a, n_b and $n_{a\overline{b}}$[1], whereas equilibrium is a function of the two parameters n_a and $n_{a\overline{b}}$. Thus, all the IMs of deviation from independence depend on the four parameters, while the IMs of deviation from equilibrium do not depend on n_b and n generally. The only exceptions to this principle are IPEE [7] and the Least Contradiction [4]. IPEE (see the formula in Appendix A) measures the statistical significance of the deviation from equilibrium. It depends on n. The Least Contradiction depends on n_b (see the formula in Appendix B). This is a hybrid IM which has a fixed value at equilibrium – as the IMs of deviation from equilibrium – but decreases with n_b – as the IMs of deviation from independence.

Comparison of the filtering capacities

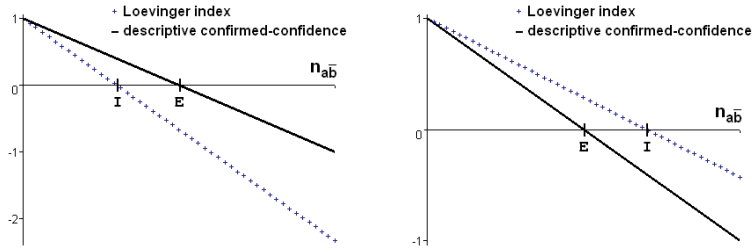
We aim at filtering the rules with a threshold on the IMs (by retaining only the high values of the IMs), and at comparing the numbers of rules that are

[1] Here we have chosen $n_{a\overline{b}}$ as a parameter, but we could have chosen another cardinality of the joint distribution of the itemsets a and b, such as n_{ab}.

rejected by the IMs of deviation from equilibrium and from independence. Let us consider a rule with the cardinalities n, n_a, n_b, and $n_{a\bar{b}}$. By varying $n_{a\bar{b}}$ with fixed n, n_a, and n_b, one can distinguish two different cases:

- Case 1: $n_b \geq \frac{n}{2}$. Then $\frac{n_a n_{\bar{b}}}{n} \leq \frac{n_a}{2}$, and the rule goes through the independence before going through the equilibrium when $n_{a\bar{b}}$ increases.
- Case 2: $n_b \leq \frac{n}{2}$. Then $\frac{n_a n_{\bar{b}}}{n} \geq \frac{n_a}{2}$, and the rule goes through the equilibrium before going through the independence when $n_{a\bar{b}}$ increases.

Let us now compare an IM of deviation from equilibrium m_{eql} and an IM of deviation from independence m_{idp} for these two cases. In order to have a fair comparison, we suppose that the two IMs have similar behaviors: same value for a logical rule, same value for equilibrium/independence, same decrease speed with regard to the negative examples. For example, m_{eql} and m_{idp} can be the Descriptive Confirmed-Confidence [18] and the Loevinger index respectively [22] (Appendix B). As shown in figure 1, m_{idp} is more filtering than m_{eql} in case 1, whereas m_{eql} is more filtering than m_{idp} in case 2. More precisely, in case 1, m_{idp} contributes to rejecting the bad rules, while in case 2 it is m_{eql}. This confirms that the IMs of deviation from equilibrium and the IMs of deviation from independence are complementary, the second ones not being systematically "better" than the first ones[2]. In particular, the IMs of deviation from equilibrium must not be neglected when itemsets are rare (low frequency). In this situation, case 2 is more frequent than case 1.



(a) case 1 ($n_b \geq \frac{n}{2}$) (b) case 2 ($n_b \leq \frac{n}{2}$)

Fig. 1. Comparison of Descriptive Confirmed-Confidence and Loevinger index (E: equilibrium, I: independence)

In our IM taxonomy, the subject of an IM could be the deviation from independence or the deviation from equilibrium. However, as some IMs do not assess any of the two deviation, a third cluster must be added ("other measures" in Tab. 1). The IMs of this cluster generally have a fixed value

[2] Numerous authors consider that a good IM must vanish at independence (principle originally proposed in [25]). This amounts to saying that IMs of deviation from independence are better than IMs of deviation from equilibrium.

only for the rules with no negative examples ($n_{a\bar{b}} = 0$) or for the rules with no examples ($n_{ab} = 0$). Most of them are similarity measures.

3.2 Nature of an IM

The objective IMs can also be classified according to their descriptive or statistical nature.

Descriptive IMs

The descriptive (or frequential) IMs do not vary with the cardinality expansion (when all the data cardinalities are increased or decreased in equal proportion). A descriptive IM m satisfies $m(n, n_a, n_b, n_{a\bar{b}}) = m(\alpha.n, \alpha.n_a, \alpha.n_b, \alpha.n_{a\bar{b}})$ for any strictly positive constant α. These IMs take the data cardinalities into account only in a relative way (by means of the frequencies $p(a)$, $p(b)$, $p(a \; and \; \bar{b})$) and not in an absolute way (by means of the cardinalities n_a, n_b, $n_{a\bar{b}}$).

Statistical IMs

The statistical IMs are those which vary with the cardinality expansion. They take into account the size of the phenomena studied. Indeed, a rule is statistically more valid when it is accessed on a large amount of data. Among the statistical IMs, one can find the probabilistic IMs, which compare the observed distribution to an expected distribution, such as the II measure presented in Appendix A.

3.3 IM taxonomy

A taxonomy according to the nature and subject of the objective IMs is given below (Tab. 1). On the column, we can see that most of the IMs are descriptive. Another observation shows that IPEE is the only one statistical IM computing the deviation from equilibrium.

4 ARQAT tool

ARQAT (Fig. 2) is an exploratory analysis tool that embeds thirty-six objective IMs studied in surveys (See Appendix B for a complete list of selected IMs).

It provides graphical views structured in five task-oriented groups: ruleset analysis, correlation and clustering analysis, interesting rules analysis, sensitivity analysis, and comparative analysis.

Subject \ Nature	Descriptive IMs	Statistical IMs
Measures of deviation from equilibrium	– Confidence (5), – Laplace (21), – Sebag & Schoenauer (31), – Example & Contra-Example (13), – Descriptive Confirm (9), – Descriptive Confirmed-Confidence (10), – Least Contradiction (22)	– IPEE (16)
Measures of deviation from independence	– Phi-Coefficient (28), – Lift (23), – Loevinger (25), – Conviction (6), – Dependency (8), – Pavillon (27), – J-measure (18), – Gini-index (14), – TIC (33), – Collective Strength (4), – Odds Ratio (26), – Yules's Q (34), – Yule's Y (35), – Klosgen (20), – Kappa (19)	– II (15), – EII$\alpha = 1$ (11), – EII$\alpha = 2$ (12), – Lerman (24), – Rule Interest (30)
Other measures	– Support (32), – Causal Support (3), – Jaccard (17), – Cosine (7), – Causal Confidence (0), – Causal Confirm (1), – Causal Confirmed-Confidence (2), – Putative Causal Dependency (29)	

Table 1. Taxonomy of the objective IMs

The ARQAT input is a set of association rules R where each association rule $a \rightarrow b$ must be associated with the four cardinalities n, n_a, n_b, and $n_{a\bar{b}}$.

In the first stage, the input ruleset is preprocessed in order to compute the IM values of each rule, and the correlations between all IM pairs. The results are stored in two tables: an IM table (R×I) where rows are rules and columns are IM values, and a correlation matrix (I×I) crossing IMs. At this stage, the ruleset may also be sampled (filtering box in Fig. 2) in order to focus the study on a more restricted subset of rules.

In the second stage, the data-analyst can drive the graphical exploration of results through a classical web-browser. ARQAT is structured in five groups of task-oriented views. The first group (1 in Fig. 2) is dedicated to ruleset and simple IM statistics to better understand the structure of the IM table (R×I). The second group (2) is oriented to the study of IM correlation in table (I×I) and IM clustering in order to select the most suitable IMs. The third one (3) focuses on rule ordering to select the most interesting rules. The fourth group (4) proposes to study the sensitivity of IMs. The last group (5) offers the possibility to compare the results obtained from different rulesets.

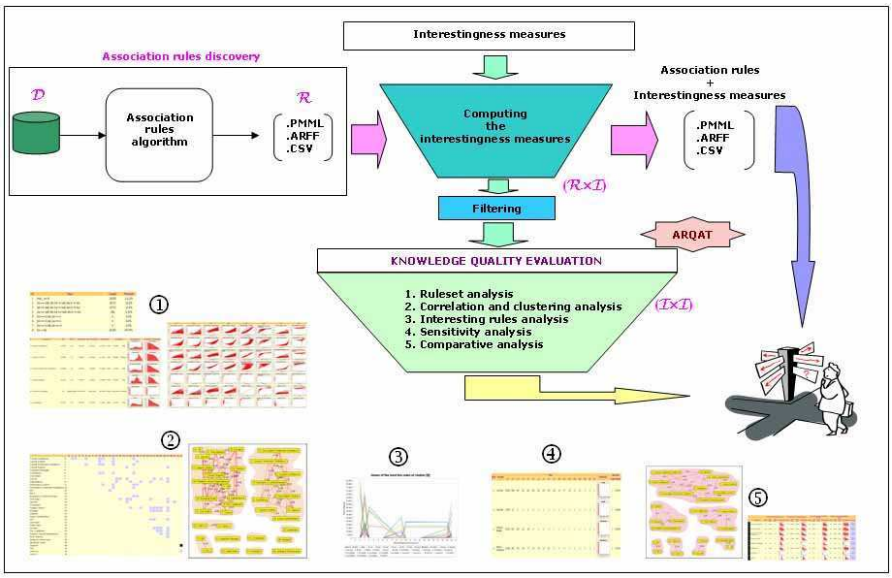


Fig. 2. ARQAT structure.

In this section, we focus on the description of the first three groups and we illustrate them with the same ruleset: 123228 association rules extracted by Apriori algorithm (support 12%) from the mushroom dataset [23].

4.1 Ruleset statistics

The basic statistics are summarized on three views of ARQAT. The first one, ruleset characteristics, shows the distributions underlying rule cardinalities, in order to detect "borderline cases". For instance, in Tab. 2, the first line gives the number of "logical" rules i.e. rules without negative examples. We can notice that the number of logical rules is here very high (≈13%).

N	Type	Count	Percent
1	$n_{a\bar{b}} = 0$	16158	13.11%
2	$n_{a\bar{b}} = 0$ & $na < nb$	15772	12.80%
3	$n_{a\bar{b}} = 0$ & $na < nb$ & $nb = n$	0	0.00%
4	$n_a > n_b$	61355	49.79%
5	$nb = n$	0	0.00%

Table 2. Some ruleset characteristics of the mushroom ruleset.

The second view, IM distribution (Fig. 3), draws the histograms for each IM. The distributions are also completed with classically statistical indexes :

minimum, maximum, average, standard deviation, skewness and kurtosis values. In Fig. 3, one can see that Confidence (line 5) has an irregular distribution and a great number of rules with 100% confidence; it is very different from Causal Confirm (line 1).

The third view, joint-distribution analysis (Fig. 4), shows the scatterplot matrix of all IM pairs. This graphical matrix is very useful to see the details of the relationships between IMs. For instance, Fig. 4 shows four disagreement shapes: Rule Interest vs Yule's Q (4), Sebag & Schoenauer vs Yule's Y (5), Support vs TIC (6), and Yule's Y vs Support (7) (highly uncorrelated). On the other hand, we can notice four agreement shapes on Phi-Coefficient vs Putative Causal Dependency (1), Phi-Coefficient vs Rule Interest (2), Putative Causal Dependency vs Rule Interest (3), and Yule's Q vs Yule's Y (8) (highly correlated).

N°	Measure	Min	Max	Average	Std. Deviation	Skewness		Kurtosis		Histogram	Inverse-cumulative
0	Causal Confidence	0.0614	1.0	0.6246	0.2712	-0.2033	Left	-1.0001	Flat		
1	Causal Confirm	-1.5951	1.0	0.1084	0.5744	-1.1203	Left	0.8805	Peaked		
2	Causal Confirmed-Confidence	-0.8169	1.0	0.2018	0.5448	0.0286	Right	-1.2036	Flat		
3	Causal Support	0.1198	1.0	0.5542	0.2051	-0.0995	Left	-0.5643	Flat		
4	Collective Strength	0.0	785856.8441	2328.2323	7944.9114	30.5993	Right	1932.4436	Peaked		
5	Confidence	0.1217	1.0	0.5772	0.2799	0.194	Right	-1.3159	Flat		
6	Conviction	0.0	219.4653	2.1238	6.3156	14.7085	Right	287.2374	Peaked		

Fig. 3. Distribution of some IMs on the mushroom dataset.

4.2 Correlation analysis

This task aims at delivering IM clustering and facilitating the choice of a subset of IMs that is best-adapted to describe the ruleset. The correlation values between IM pairs are computed in the preprocessing stage by using the Pearson's correlation coefficient and stored in the correlation matrix (I×I). Two visual representations are proposed. The first one is a simple summary matrix in which each significant correlation value is visually associated with a

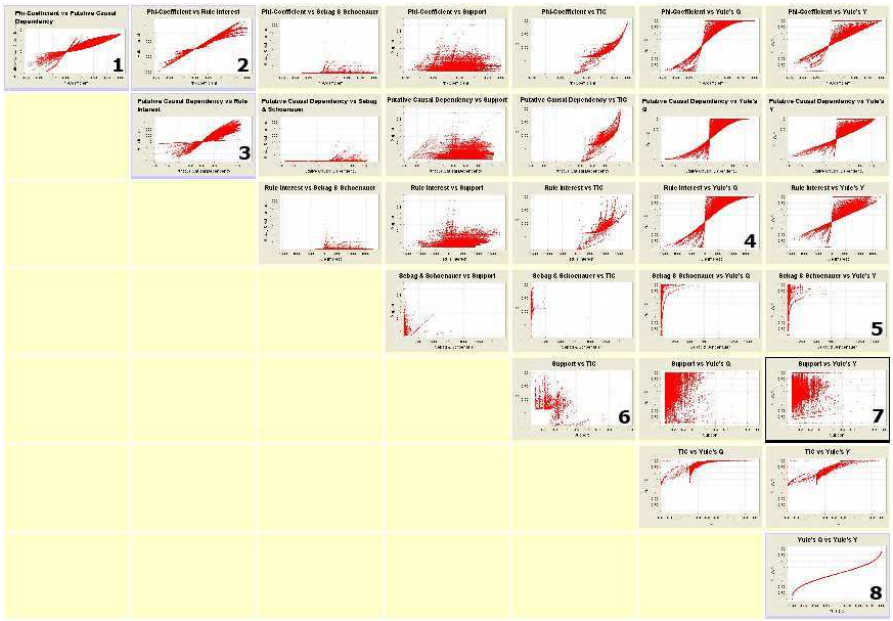


Fig. 4. Scatterplot matrix of joint-distributions on the mushroom dataset.

different color (a level of gray). For instance, the furthest right dark cell from Fig. 5 shows a low correlation value between Yule's Y and Support. The other seventy-nine gray cells correspond to high correlation values.

The second one (Fig. 6) is a graph-based view of the correlation matrix. As graphs are a good means to offer relevant visual insights on data structure, the correlation matrix is used as the relation of an undirected and valued graph, called "correlation graph". In a correlation graph, a vertex represents an IM and an edge value is the correlation value between two vertices/IMs. We also add The possibility to set a minimal threshold τ (maximal threshold θ respectively) to retain only the edges associated with a high correlation (respectively low correlation); the associated subgraphs are denoted by CG+ and CG0.

These two subgraphs can then be processed in order to extract clusters of IMs: each cluster is defined as a connected subgraph. In CG+, each cluster gathers correlated or anti-correlated IMs that may be interpreted similarly: they deliver a close point of view on data. Moreover, in CG0, each cluster contains uncorrelated IMs: i.e. IMs that deliver a different point of view.

Hence, as each graph depends on a specific ruleset, the user can use the graphs as data insight, which graphically help him/her select the minimal set of the IMs best adapted to his/her data. For instance in Fig. 6, CG+ graph contains twelve clusters on thirty-six IMs. The user can select the most representative IM in each cluster, and then retain it to validate the rules.

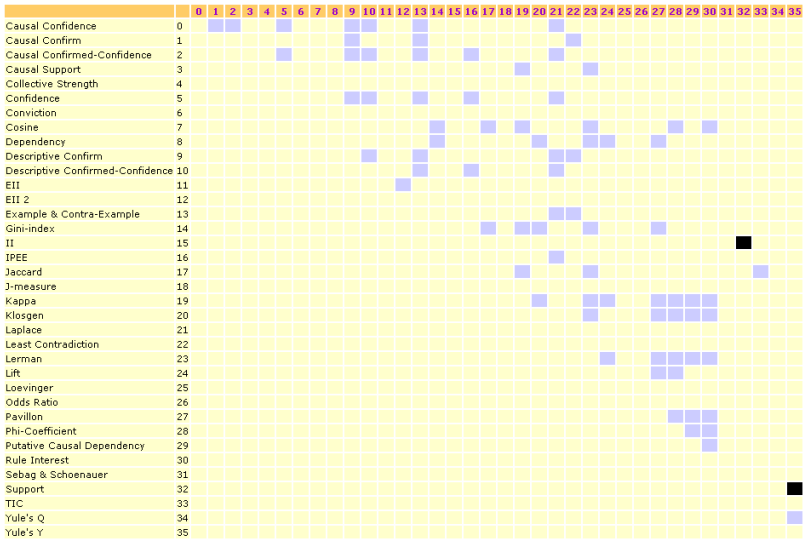


Fig. 5. Summary matrix of correlations on the mushroom dataset.

A close observation on the CG0 graph (Fig. 6) shows an uncorrelated cluster formed by II, Support and Yule's Y measures (also the two dark cells in Fig. 5). This observation is confirmed on Fig. 4 (7). CG+ graph shows a trivial cluster where Yule's Q and Yule's Y are strongly correlated. This is also confirmed in Fig. 4 (8), showing a functional dependency between the two IMs. These two examples show the interest of using the scatterplot matrix complementarily (Fig. 4) with the correlation graphs CG0, CG+ (Fig. 6) in order to evaluate the nature of the correlation links, and overcome the limits of the correlation coefficient.

4.3 Interesting rule analysis

In order to help a user select the most interesting rules, two specific views are implemented. The first view (Fig. 7) collects a set of a given number of interesting rules for each IM in one cluster, in order to answer the question: how interesting are the rules of this cluster?. The selected rules can alternatively be visualized with parallel coordinate drawing (Fig. 8). The main interest of such a drawing is to rapidly see the IM rankings of the rules.

These two views can be used with the IM values of a rule or alternatively with the rank of the value. For instance, Fig. 7 and Fig. 8 use the rank to evaluate the union of the ten interesting rules for each of the ten IMs in the C0 cluster (see Fig. 6). The Y-axis in Fig. 8 holds the rule rank for the corresponding IM. By observing the concentration lines on low rank values, one can obtain four IMs: Confidence(5), Descriptive Confirmed-Confidence(10),

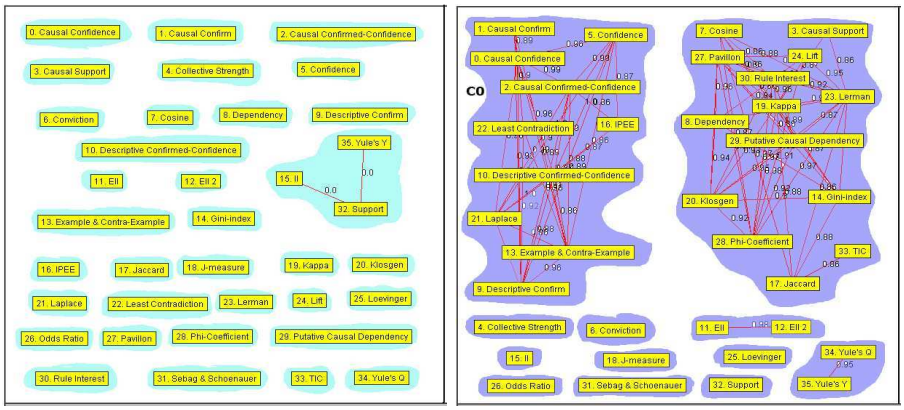


Fig. 6. CG0 and CG+ graphs on the mushroom dataset (clusters are highlighted with a gray background).

Example & Contra-Example(13), and IPEE (16) (on points 1, 2, 3, 4 respectively) that are good for a majority of interesting rules. This can also be retrieved from columns 5, 10, 13, 16 of Fig. 7. Among these four IMs, IPEE is the most suitable choice because of the lowest rule ranks obtained.

Measure Order	0	1	2	5	9	10	13	21	22	16	Rule's presentation
30	R107560	1	19121	1	1	41	1	1	8	5388	1	BROAD FREE ONE ==>veil_color=WHITE
31	R107562	1	18997	1	1	41	1	1	8	5361	1	BROAD ONE veil_color=WHITE ==>FREE
32	R107594	1	8972	1	1	18	1	1	3	2574	1	CLOSE FREE ONE ==>veil_color=WHITE
33	R107596	1	8914	1	1	18	1	1	3	2564	1	CLOSE ONE veil_color=WHITE ==>FREE
34	R122275	1	13800	1	1	32	1	1	5	3977	1	BROAD FREE ==>veil_color=WHITE
35	R122283	1	18299	1	1	38	1	1	6	5145	1	FREE stalk_surf_above=SMOOTH ==>veil_color=WHITE
36	R122285	1	18179	1	1	38	1	1	6	5134	1	stalk_surf_above=SMOOTH veil_color=WHITE ==>FREE
37	R122296	1	20903	1	1	55	1	1	10	6193	1	FREE stalk_surf_below=SMOOTH ==>veil_color=WHITE
38	R122308	65969	8772	40612	23743	10	23743	23743	23714	1013	1	FREE ==>ONE veil_color=WHITE

Fig. 7. Union of the ten interesting rules of the cluster $C0$ on the mushroom dataset (extract).

5 Focus on graph-based clustering approach

When considering a large set of IMs, the graph-based view of the correlation matrix may be quite complex. In order to highlight the more "natural" clusters, we propose to construct two types of subgraphs : the correlated ($CG+$) and the uncorrelated ($CG0$) partial subgraph. In this section we present the different filtering thresholds for their construction. We also extend the correlation graphs to graphs of stable clusters ($\overline{CG0}$ and $\overline{CG+}$) in order to compare several rulesets.

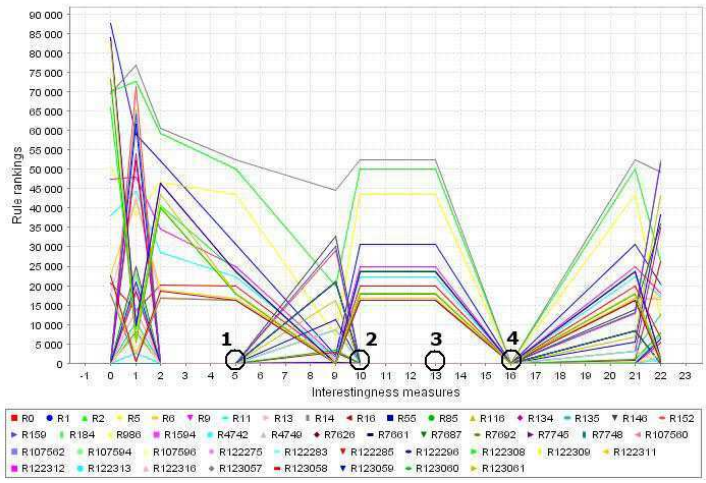


Fig. 8. Plot of the union of the ten interesting rules of the cluster $C0$ on the mushroom dataset.

5.1 Principles

Let $R(D) = \{r_1, r_2, ..., r_p\}$ denote a set of p association rules derived from a dataset D. Each rule $a \rightarrow b$ is described by its itemsets (a, b) and its cardinalities $(n, n_a, n_b, n_{a\bar{b}})$. Let M be the set of q available IMs for our analysis $M = \{m_1, m_2, ..., m_q\}$. Each IM is a numerical function on rule cardinalities: $m(a \rightarrow b) = f(n, n_a, n_b, n_{a\bar{b}})$. For each IM $m_i \in M$, we can construct a vector $m_i(R) = \{m_{i1}, m_{i2}, ..., m_{ip}\}, i = 1..q$, where m_{ij} corresponds to the calculated value of the IM m_i for a given rule r_j.

The correlation value between any two IMs $m_i, m_j\{i, j = 1..q\}$ on the set of rules R is calculated by using a Pearson's correlation coefficient $\rho(m_i, m_j)$ [27], where $\overline{m_i}, \overline{m_j}$ are the average values calculated of vector $m_i(R)$ and $m_j(R)$ respectively:

$$\rho(m_i, m_j) = \frac{\sum_{k=1}^{p}[(m_{ik} - \overline{m_i})(m_{jk} - \overline{m_j})]}{\sqrt{[\sum_{k=1}^{p}(m_{ik} - \overline{m_i})^2][\sum_{k=1}^{p}(m_{jk} - \overline{m_j})^2]}}$$

In order to make the interpretation of the large set of correlation values easier, we introduce the following definitions:

Definition 3. Two IMs m_i and m_j are *τ-correlated* with respect to the dataset D if their absolute correlation value is greater than or equal to a given threshold τ: $|\rho(m_i, m_j)| \geq \tau$. And, conversely, two IMs m_i and m_j are *θ-uncorrelated* with respect to the dataset D if the absolute value of their correlation value is lower than or equal to a threshold value θ: $|\rho(m_i, m_j)| \leq \theta$.

For θ-uncorrelated IMs, we use a statistical test of significance by choosing a level of significance of the test $\alpha = 0.05$ for hypothesis testing (common values for α are: $\alpha = 0.1, 0.05, 0.005$). The threshold θ is then calculated by the following formula: $\theta = 1.960/\sqrt{p}$ in a population of size p [27]. The assignment $\tau = 0.85$ of τ-correlated is used because this value is widely acceptable in the literature.

As the correlation coefficient is symmetrical, the $q(q-1)/2$ correlation values can be stored in one half of the table $q \times q$. This table $(I \times I)$ can also be viewed as the relation of an undirected and valued graph called correlation graph, in which a vertex value is an IM and an edge value is the correlation value between two vertices/IMs.

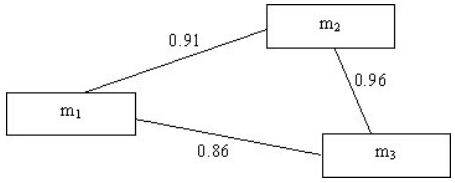


Fig. 9. An illustration of the correlation graph.

For instance, Fig. 9 can be the correlation graph obtained on five association rules $R(D) = \{r_1, r_2, r_3, r_4, r_5\}$ extracted from a dataset D and three IMs $M = \{m_1, m_2, m_3\}$ whose values and correlations are given in Tab. 3.

$R \times I$	m_1	m_2	m_3
r_1	0.84	0.89	0.91
r_2	0.86	0.90	0.93
r_3	0.88	0.94	0.97
r_4	0.94	0.95	0.99
r_5	0.83	0.87	0.84

$I \times I$	m_1	m_2	m_3
m_1		0.91	0.86
m_2			0.96
m_3			

Table 3. Correlation values for three IMs and five association rules.

5.2 Correlated versus uncorrelated graphs

Unfortunately, when the correlation graph is complete, it is not directly human-readable. We need to define two transformations in order to extract more limited and readable subgraphs. By using definition 3, we can extract the *correlated partial subgraph (CG+)*: the subgraph composed of edges associated with a τ-correlated. On the same way, the *uncorrelated partial subgraph (CG0)* where we only retain edges associated with correlation values close to 0 (θ-uncorrelated).

These two partial subgraphs can then be used as a visualization support in order to observe the correlative liaisons between IMs.

We can also observe the clusters of IMs corresponding with the connected parts of the graphs.

5.3 Extension to graph of stable clusters

In order to facilitate the comparison between several correlation matrices, we have introduced some extensions to define the stable clusters between IMs.

Definition 4. The $\overline{CG+}$ graph (respectively $\overline{CG0}$ graph) of a set of k rulesets $R = \{R(D_1), ..., R(D_k)\}$ is defined as the average graph of intersection of the k partially correlated (respectively uncorrelated) subgraphs $CG+_k$ (respectively $CG0_k$) calculated on R. Hence, each edge of $\overline{CG+}$ (respectively $\overline{CG0}$) is associated with the average value of the corresponding edge in the k $CG+_k$ graphs. Therefore, the $\overline{CG+}$ (respectively $\overline{CG0}$) graph allows visualizing the strongly (respectively weakly) stable correlations, as being common to k studied rulesets.

Definition 5. We call τ-stable (respectively θ-stable) clusters the connected part of the $\overline{CG+}$ (respectively $\overline{CG0}$) graph.

6 Study of IM behavior on two prototypical and opposite datasets

We have applied our method to two "opposite" datasets: D_1 and D_2, in order to compare correlation behavior and more precisely, to discover some stable clusters.

6.1 Data description

Our experiments are based on the categorical mushroom dataset (D_1) from Irvine machine-learning database repository and a synthetic dataset (D_2). The latter is obtained by simulating the transactions of customers in retail businesses, the dataset was generated using the IBM synthetic data generator [3]. D_2 has the typical characteristic of the Agrawal dataset T5.I2.D10k. We also generate the set of association rules (ruleset) R_1 (respectively R_2) from the dataset D_1 (respectively D_2) using the Apriori algorithm [2] [3]. For a closer evaluation of the IM behavior of the most interesting rules from these two rulesets, we have extracted $R_1^{'}$ (respectively $R_2^{'}$) from R_1 (respectively R_2) as the union of the first 1000 rules ($\approx 1\%$, ordered by decreasing IM values) issued from each IM (see Tab. 4).

In our experiment, we compared and analyzed the thirty-six IMs defined in Appendix B. We must notice that EII($\alpha = 1$) and EII($\alpha = 2$) are two entropic versions of the II measure.

Dataset	Items (Average length)	Transactions	Number of rules (support threshold)	$R(D)$	θ	τ	$R(D)$
D_1	118 (22)	8416	123228 (12%)	R_1	0.005	0.85	R_1
			10431 (12%)	$R_1^{'}$	0.020	0.85	$R_1^{'}$
D_2	81 (5)	9650	102808 (0.093%)	R_2	0.003	0.85	R_2
			7452 (0.093%)	$R_2^{'}$	0.012	0.85	$R_2^{'}$

Table 4. Description of the datasets.

6.2 Discussion

The analysis aims at finding stable relations between the IMs studied over the four rulesets. We investigate in: (1) the $\overline{CG0}$ graphs in order to identify the IMs that do not agree for ranking the rules, (2) the $\overline{CG+}$ graph in order to find the IMs that do agree for ranking the rules.

Ruleset	Number of correlations		Number of clusters	
	τ-correlated	θ-uncorrelated	CG+	CG0
R_1	79	2	12	34
$R_1^{'}$	91	15	12	21
R_2	65	0	14	36
$R_2^{'}$	67	17	12	20

Table 5. Comparison of correlation.

CG+ and *CG0*

Fig. 10 shows four $CG+$ graphs obtained from the four rulesets. As seen before, the sample rulesets and the original rulesets have close results so we can use the sample rulesets for representing the original rulesets. This observation is useful when we evaluate the CG+ graphs but not for CG0 graphs. For example, with the CG+ graph of R_1 (Fig. 10), one can choose the largest cluster containing the fourteen IMs (Causal Support, Pavillon, Lift, Lerman, Putative Causal Dependency, Rule Interest, Phi-Coefficient, Klosgen, Dependency, Kappa, Gini-index, Cosine, Jaccard, TIC) for his/her first choice. In this cluster one can also see the weak relation between TIC and the other IMs of the cluster. Tab. 5 also shows the two opposite tendencies obtained from the number of τ-correlated computed: $79(R_1) \rightarrow 91(R_1^{'}), 65(R_2) \rightarrow 67(R_2^{'})$.

With the four CG0 graphs (Fig. 11), one can easily see that the number of θ-uncorrelated increases when the most interesting rules are selected: $2(R_1) \rightarrow 15(R_1^{'}), 0(R_2) \rightarrow 17(R_2^{'})$ (Fig. 11, Tab. 5).

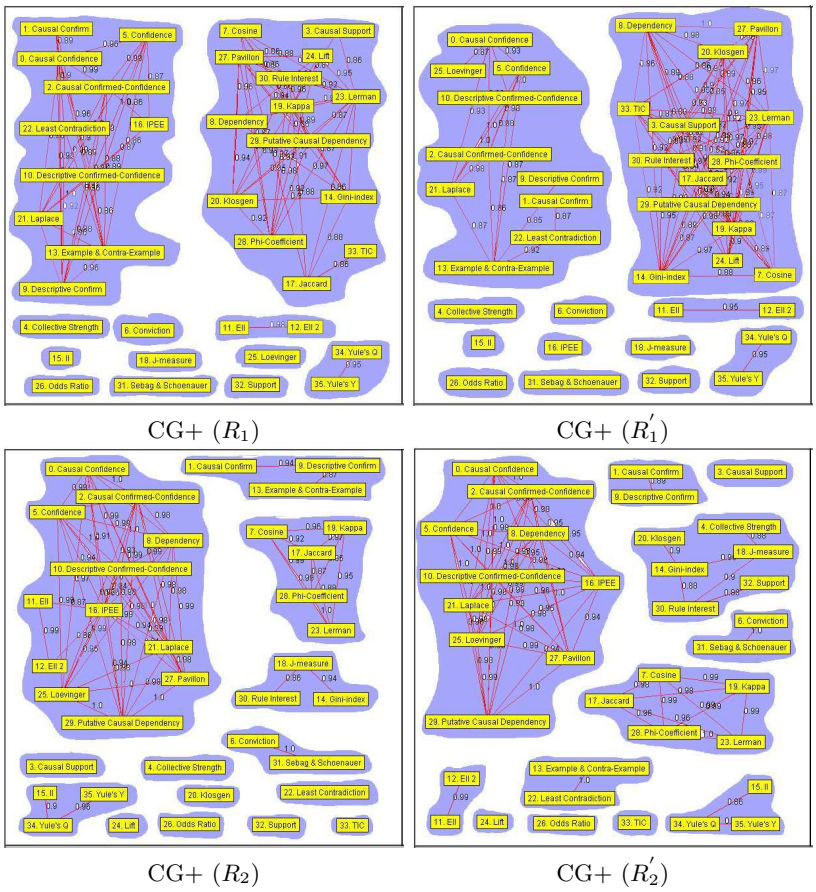


Fig. 10. CG+ graphs (clusters are highlighted in gray).

$\overline{CG0}$ graphs: uncorrelated stability

Uncorrelated graphs first show that there are no θ-stable clusters that appear on the four rulesets studied in Fig. 11. Secondly, there is no $\overline{CG0}$ graph from these datasets. A close observation of four CG0 graphs shows that at least one IM in each cluster will later be clustered around in a τ-stable cluster of $\overline{CG+}$ graph (Fig. 11, Fig. 12) like Yule's Y, Putative Causal Dependency, EII($\alpha = 2$), Cosine, Laplace so that the stronger the θ-uncorrelated, the more interesting the IM that participated in the θ-uncorrelated.

$\overline{CG+}$ graph: correlated stability

From Tab. 5, we can see that, R_1' is approximately twice as correlated as R_2'. As seen in Fig. 12, five τ-stable clusters found come from the datasets studied.

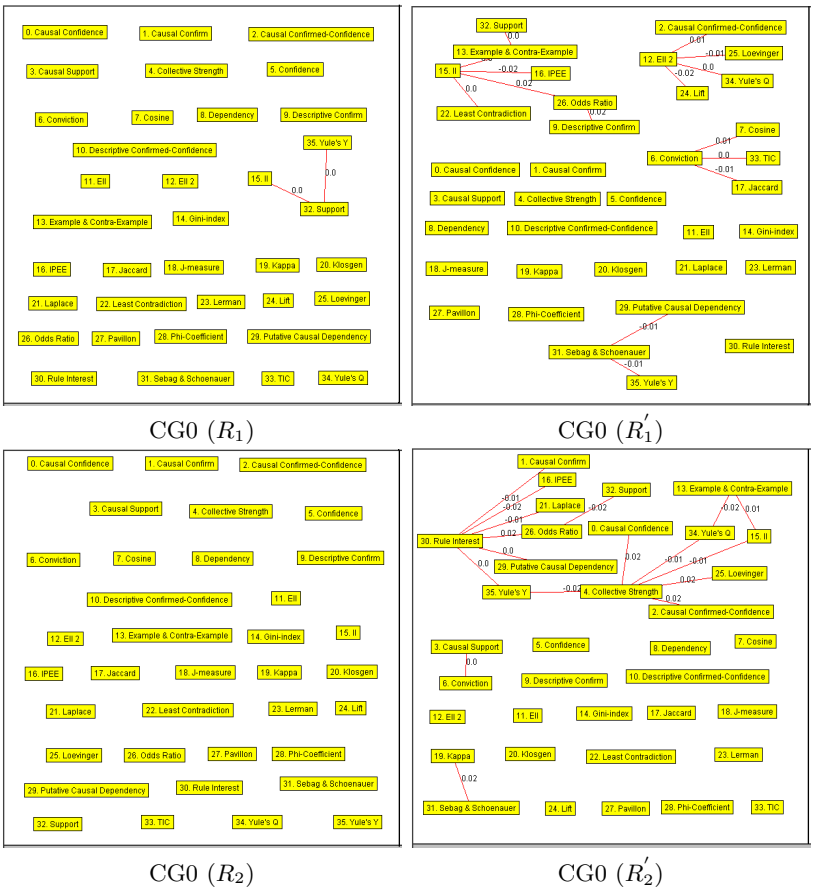


Fig. 11. CG0 graphs.

By briefly analyzing these τ-stable clusters, some interesting observations are drawn.

(C1), the largest cluster, (Confidence, Causal Confidence, Causal Confirmed-Confidence, Descriptive Confirmed-Confidence, Laplace) has most of its IMs extended from Confidence measure. From this cluster, we can easily see a highly connected component – each vertex must have an edge with the other vertices – indicating the strong agreement of the five IMs.

According to the taxonomy (Tab. 1), this cluster is associated with descriptive IMs that are sensible to equilibrium.

(C2), another cluster, has two highly connected components which are formed by Phi-Coefficient, Lerman, Kappa, Cosine and Jaccard. Most of these IMs are similarity measures. According to the taxonomy (Tab. 1) this cluster is to measure the deviation from independence.

(C3), this cluster (Dependency, Pavillon, Putative Causal Dependency) is interesting because almost all the IMs of this cluster are reasonably well correlated. The nature of these IMs are descriptive.

(C4), is a cluster formed by EII and EII 2, which are two IMs obtained with different parameters of the same original formula. This cluster has many extended directions to evaluate the entropy of II.

(C5), Yule's Q and Yule's Y, brings out a trivial observation because these IMs are derived from Odds Ratio measure. Both IMs are descriptive and measuring of deviation from independence.

In looking for τ-stable clusters, we have found the τ-correlated that exist between various IMs and we have identified five τ-stable clusters. Each τ-stable cluster forms a subgraph in a $\overline{CG+}$ graph, also contains a highly connected component. Therefore, we can choose a representative IM for each cluster. For example, in our experiment, we have five representative IMs for all the thirty-six IMs. How we can choose a representative IM is also an interesting study for the future. In the first approach, we can select the IM that has the highest number of relations with the others: Causal Confidence, Cosine, Klosgen, EII($\alpha = 2$), and Yule's Y. The stronger the τ-stable cluster, the more interesting the representative IM. An important observation is that, the existence of highly connected graphs represents a strong agreement with a τ-stable cluster. We have reached significant information: *τ-stable clusters can be obtained from different IMs and rulesets.* The different IMs imply taking into account both their mathematical definitions and their respective significance. The datasets are both highly correlated and lowly correlated.

7 Conclusion

We have studied and compared the various IMs described in the literature in order to help the decision-maker to better understand the behavior of the IMs in the stage of post-processing of association rules. A new approach called correlation graph implemented by a new tool, ARQAT, with two types: CG+ and CG0 is proposed to evaluate IMs by using graphs as a visual insight on the data.

With this approach, the decision-maker has a few IMs to decide and as a graphical representation to select the most interesting rules to examine. Another interesting result obtained from this work is that we have found some stable clusters between IMs, five such τ-stable clusters have been found with the $\overline{CG+}$ graph. Our approach is highly related to the real value of the dataset and the number of proposed IMs.

Our future research will investigate the two following directions: first, we will improve the correlation analysis by introducing a better measure than linear correlation whose limits are stressed in the literature; second, we will also improve the IM clustering analysis with IM aggregation techniques to facilitate the user's decision making from the most suitable IMs.

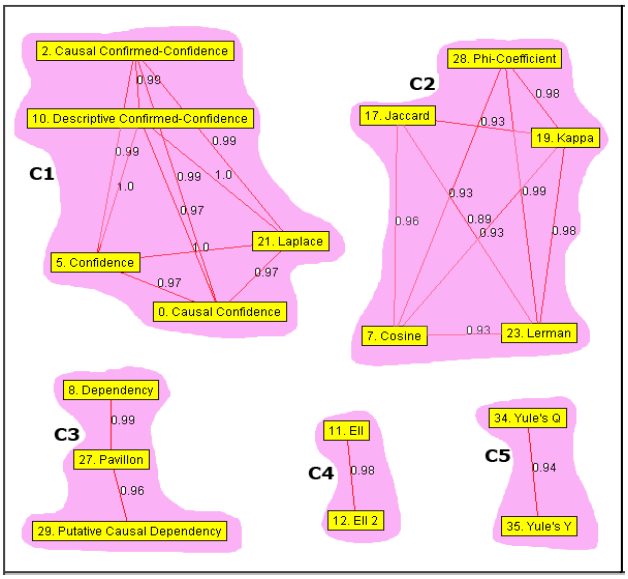


Fig. 12. $\overline{CG+}$ graph.

References

1. Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases. Proceedings of the ACM-SIGMOD International Conference on Management of Data. Washington DC, USA (1993) 207–216
2. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. VLDB'94, Proceedings of the 20th International Conference on Very Large Data Bases. Santiago, Chile (1994) 487–499
3. Agrawal, R., Mannila, H., Srikant, R., Toivonen H., Verkano, A.I.: Fast discovery of association rules. Advances in Knowledge Discovery in Databases. (1996) 307–328
4. Azé, J., Kodratoff, Y.: A study of the Effect of Noisy Data in Rule Extraction Systems. EMCSR'02, Proceedings of the Sixteenth European Meeting on Cybernetics and Systems Research. (2002) 781–786
5. Bayardo, Jr.R.J., Agrawal, R.: Mining the most interesting rules. KDD'99, Proceedings of the Fifth ACM SIGKDD international conference on Knowledge discovery and data mining. San Diego, CA, USA (1999) 145–154
6. Blanchard, J., Guillet, F., Gras, R., and Briand, H.: Using information-theoretic measures to assess association rule interestingness. ICDM'05, Proceedings of the 5th IEEE International Conference on Data Mining, IEEE Computer Society Press, (2005) 66–73.
7. Blanchard, J., Guillet, F., Gras, R., Briand, H.: Assessing rule interestingness with a probabilistic measure of deviation from equilibrium. ASMDA'05, Proceedings of the 11th International Symposium on Applied Stochastic Models and Data Analysis. (2005) 191–200

8. Blanchard, J., Guillet, F., Gras, R., Briand, H.: Mesurer la qualité des règles et de leurs contraposées avec le taux informationnel TIC. EGC'04, Actes de 4èmes journées d'Extraction et de Gestion des Connaissances, RNTI-E-2, Vol. 1. Cépaduès Editions, Clermont Ferrand, France (2004) 287–298 (in French)
9. Blanchard, J., Kuntz, P., Guillet, F., Gras, R.: Implication Intensity: from the basic statistical definition to the entropic version. Statistical Data Mining and Knowledge Discovery, Chapter 28. Chapman & Hall, CRC Press (2003) 475–493
10. Freitas, A.A.: On rule interestingness measures. Knowledge-Based Systems, 12(5-6). (1999) 309–315
11. Gavrilov, M., Anguelov, D., Indyk, P., and Motwani, R.: Mining the stock market: which measure is best?. KDD'00, Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining . Boston, MA, USA (2000) 487–496.
12. Gras, R., Couturier, R., Blanchard, J., Briand, H., Kuntz, P., Peter, P.: Quelques critères pour une mesure de qualité de règles d'association. Mesures de Qualité pour la Fouille de Données, RNTI-E-1. Cépaduès Editions (2004) 3–31 (in French)
13. Gras, R.: L'implication statistique - Nouvelle méthode exploratoire de données. La Pensée Sauvage Édition (1996) (in French)
14. Hilderman, R.J., Hamilton, H.J.: Knowledge Discovery and Measures of Interestingness. Kluwer Academic Publishers (2001)
15. Klemettinen, M., Mannila, H., Ronkainen, P., Toivonen, H., and Verkano, A.I.: Finding interesting rules from larges sets of discovered association rules. ICIKM'94, Proceedings of the Third International Conference on Information and Knowledge Management. Ed. Nabil R. Adam, Bharat K. Bhargava and Yelena Yesha, Gaithersburg, Maryland. ACM Press, (1994) 401–407.
16. Huynh, X.-H., Guillet, F., Briand, H.: Clustering interestingness measures with positive correlation. ICEIS'05, Proceedings of the 7th International Conference on Enterprise Information Systems. (2005) 248–253
17. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York, (1990)
18. Kodratoff, Y.: Comparing Machine Learning and Knowledge Discovery in DataBases: An Application to Knowledge Discovery in Texts. Machine Learning and Its Applications, LNCS 2049. Springer-Verlag, (2001) 1–21
19. Kononenco, I.: On biases in estimating multi-valued attributes. IJCAI'95. (1995) 1034–1040
20. Lenca, P., Lallich, S., Vaillant, B.: On the robustness of association rules. Proceedings of the IEEE International Conference on Cybernetics and Intelligent Systems. (2006) 596–601
21. Liu, B., Hsu, W., Mun, L., Lee, H.: Finding interestingness patterns using user expectations. IEEE Transactions on Knowledge and Data Mining (11). (1999) 817–832
22. Loevinger, J.: A systematic approach to the construction and evaluation of tests of ability. Psychological Monographs. (1947)
23. Newman, D.J., Hettich,S., Blake, C.L., Merz, C.J.: [UCI] Repository of machine learning databases, http://www.ics.uci.edu/∼mlearn/MLRepository.html. University of California, Irvine, Department of Information and Computer Sciences, (1998).
24. Padmanabhan, B., Tuzhilin, A. : A belief-driven method for discovering unexpected patterns. KDD'98, Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining. (1998) 94–100

25. Piatetsky-Shapiro, G.: Discovery, analysis and presentation of strong rules. Knowledge Discovery in Databases, G. Piatetsky-Shapiro and W. Frawley editors. MIT Press, Cambridge, MA (1991) 229–248
26. Piatetsky-Shapiro, G., Steingold, S.: Measuring Lift Quality in Database Marketing. SIGKDD Explorations 2(2). (2000) 76–80
27. Ross, S.M.: Introduction to probability and statistics for engineers and scientists. Wiley, (1987)
28. Sebag, M., Schoenauer, M.: Generation of rules with certainty and confidence factors from incomplete and incoherent learning bases. EKAW'88, Proceedings of the European Knowledge Acquisition Workshop. Gesellschaft fr Mathematik und Datenverarbeitung mbH (1988) 28.1–28.20
29. Silberschatz, A., Tuzhilin, A.: What makes patterns interesting in knowledge discovery systems. IEEE Transactions on Knowledge Data Engineering 8(6). (1996) 970–974
30. Tan, P.N., Kumar, V., Srivastava, J.: Selecting the right objective measure for association analysis. Information Systems 29(4). (2004) 293–313
31. Vaillant, B., Lenca, P., Lallich, S.: A clustering of interestingness measures. DS'04, the 7th International Conference on Discovery Science LNAI 3245. (2004) 290–297
32. Vaillant, B., Lallich, S., Lenca, P.: Modeling of the counter-examples and association rules interestingness measures behavior. The 2006 International Conference on Data Mining. (2006)
33. Zhao, Y., Karypis, G.: Criterion functions for document clustering: experiments and analysis. Technical Report TR01-40, Deparment of Computer Science, University of Minnesota. (2001) 1–30

A Complementary IMs: II, EII, TIC, IPEE

A.1 Implication Intensity

Initially introduced by Gras [13], the implicative intensity II aims at quantifying the "surprisingness" of a rule.

Intuitively, it is more surprising to discover that a rule has a small number of negative examples when the dataset is large. Hence, the objective of the implicative intensity is to express the unlikelihood of $n_{a\bar{b}}$ in T.

More precisely, we compare the observed number of negative examples $n_{a\bar{b}}$ with the number $N_{a\bar{b}}$ of expected negative examples for an independence hypothesis. Let us assume that we randomly draw two subsets U and V in T with respectively n_a and n_b transactions. Then, $N_{a\bar{b}} = \left|U \bigcap \overline{V}\right|$ is the random variable associated with the number of negative examples in this random model.

Definition 6. The implicative intensity II of the rule $a \rightarrow b$ is defined by

$$II\left(a \rightarrow b\right) = 1 - p\left(N_{a\bar{b}} \leq n_{a\bar{b}}\right)$$

if $n_a \neq n$; otherwise

$$II\left(a \rightarrow b\right) = 0$$

.

In practice, the distribution of $N_{a\overline{b}}$ depends on the random drawing pattern. We here consider a hyper-geometric law: $p\left(N_{a\overline{b}} = k\right) = \frac{C_{n\overline{b}}^{n\overline{a}-k} C_{n_a}^{k}}{C_n^{n_b}}$. The effective value of II can be easily computed with this recursive formula. Other models based on the binomial law and the Poisson distribution have been proposed.

A.2 Entropic Implication Intensity

Definition 6 essentially measures the surprisingness of the rule $a \rightarrow b$. However, taking the contrapositive $\overline{b} \rightarrow \overline{a}$ into account could reinforce the assertion of the implication between a and b. Moreover, it could improve the quality of discrimination of II when the transaction set T increases: if A and B are small compared to T, their complementary sets are large and vice-versa.

For these reasons, we have introduced a weighted version of the implication intensity $\left(E\left(a,b\right).II\left(a \rightarrow b\right)\right)^{1/2}$ where $E\left(a,b\right)$ measures the disequilibrium between n_{ab} and $n_{a\overline{b}}$ – associated with $a \rightarrow b$ –, and the disequilibrium between $n_{a\overline{b}}$ and $n_{\overline{a}\overline{b}}$ – associated with its contrapositive – [9]. Intuitively, the surprise must be softened (respectively confirmed) when the number of negative examples $n_{a\overline{b}}$ is high (respectively small) for the rule and its contrapositive considering the observed cardinalities n_a and $n_{\overline{b}}$.

A well-known index for taking the cardinalities into account non-linearly is the Shannon conditional entropy. The conditional entropy $H_{b/a}$ of cases (a and b) and (a and $\overline{b}$) given a is defined by

$$H_{b/a} = -\frac{n_{ab}}{n_a} log_2 \frac{n_{ab}}{n_a} - \frac{n_{ab}}{n_a} log_2 \frac{n_{a\overline{b}}}{n_a}$$

and, similarly, we obtain the conditional entropy $H_{\overline{a}/\overline{b}}$ of cases ($\overline{a}$ and $\overline{b}$) and (a and $\overline{b}$) given $\overline{b}$. The complements of 1 for these uncertainties $1 - H$ can be interpreted as the average information collected by the realization of these experiments; the higher this information, the stronger the quality of the implication and its contrapositive.

The expected behavior of the weighted version of II is determined in three stages: *(i)* a slow reaction to the first negative examples (robustness to noise), *(ii)* an acceleration of the rejection in the neighborhood of the equilibrium, *(iii)* an increasing rejection beyond the equilibrium. The adjustment of $1 - H$ proposed in definition 6 satisfies these requirements.

Definition 7. Let $\alpha > 1$ be a fixed number. The disequilibriums are measured by $E\left(a,b\right)$, is defined by

$$E\left(a,b\right)=\left(\left(1-H_{b/a}\right)^{\alpha}.\left(1-H_{\overline{a}/\overline{b}}\right)^{\alpha}\right)^{1/2\alpha}$$

if $\frac{n_{a\overline{b}}}{n}\in\left[0,\frac{n_a}{2n}\right[\bigcap\left[0,\frac{n_b}{2n}\right[$;

$$E(a,b)=0$$

otherwise.

And, the weighted version of the implication intensity – called the entropic implication intensity – is given by

$$EII\left(a\rightarrow b\right)=\left(E\left(a,b\right).II\left(a\rightarrow b\right)\right)^{1/2}$$

Raising the conditional entropies to the power α reinforces the contrast between the different stages presented above.

A.3 TIC

In [6], we introduced DIR (Directed Information Ratio), a new rule IM which is based on information theory. DIR is the entropy decrease rate of the consequent due to the truth of the antecedent, but it is not calculated with a classical entropy function. We use an asymmetric entropy function which considers that the uncertainty is maximal (entropy = 1) when the studied modality is not the more likely. This allows DIR to differentiate two opposite rules $a\rightarrow b$ and $a\rightarrow\overline{b}$, which is not possible with the other information-theoretic measures of rule interestingness. Moreover, to our knowledge, DIR is the only rule IM which rejects both independence and equilibrium, i.e. it discards both the rules whose antecedent and consequent are negatively correlated, and the rules which have more negative examples than examples.

In [8], we proposed another IM, derived from DIR, which assesses the rules by taking their contrapositives into account. This new IM called TIC (*Taux Informationnel modulé par la Contraposée, in French*) is the geometric mean of the values of DIR for a rule and its contrapositive (if one of the two values of DIR is negative, then TIC is worth zero). Considering both the rule and its contrapositive allows to discover rules that are closer to logical implication.

A.4 IPEE

As there was no statistical IMs evaluating the deviation from equilibrium, we proposed the new measure IPEE in [7]. Following II, IPEE is based on a probabilistic model. However, while II evaluates the statistical significance of the deviation from independence, IPEE evaluates the statistical significance of the deviation from equilibrium.

B Formulas of IMs

N	Interestingness measure	$f(n, n_a, n_b, n_{a\bar{b}})$	Reference
0	Causal Confidence	$1 - \frac{1}{2}(\frac{1}{n_a} + \frac{1}{n_{\bar{b}}})n_{a\bar{b}}$	[18]
1	Causal Confirm	$\frac{n_a + n_{\bar{b}} - 4n_{a\bar{b}}}{n}$	[18]
2	Causal Confirmed-Confidence	$1 - \frac{1}{2}(\frac{3}{n_a} + \frac{1}{n_{\bar{b}}})n_{a\bar{b}}$	[18]
3	Causal Support	$\frac{n_a + n_{\bar{b}} - 2n_{a\bar{b}}}{n}$	[18]
4	Collective Strength	$\frac{(n_a - n_{a\bar{b}})(n_{\bar{b}} - n_{a\bar{b}})(n_a n_{\bar{b}} + n_b n_{\bar{a}})}{(n_a n_b + n_{\bar{a}} n_{\bar{b}})(n_b - n_a + 2n_{a\bar{b}})}$	[30]
5	Confidence	$1 - \frac{n_{a\bar{b}}}{n_a}$	[2]
6	Conviction	$\frac{n_a n_{\bar{b}}}{n n_{a\bar{b}}}$	[30]
7	Cosine	$\frac{n_a - n_{a\bar{b}}}{\sqrt{n_a n_b}}$	[30]
8	Dependency	$\lvert \frac{n_{\bar{b}}}{n} - \frac{n_{a\bar{b}}}{n_a} \rvert$	[18]
9	Descriptive Confirm	$\frac{n_a - 2n_{a\bar{b}}}{n}$	[18]
10	Descriptive Confirmed-Confidence	$1 - 2\frac{n_{a\bar{b}}}{n_a}$	[18]
11	EII ($\alpha = 1$)	$\sqrt{\varphi \times I^{\frac{1}{2\alpha}}}$	[9]
12	EII ($\alpha = 2$)	$\sqrt{\varphi \times I^{\frac{1}{2\alpha}}}$	[9]
13	Example & Contra-Example	$1 - \frac{n_{a\bar{b}}}{n_a - n_{a\bar{b}}}$	[13]
14	Gini-index	$\frac{(n_a - n_{a\bar{b}})^2 + n_{a\bar{b}}^2}{n n_a} + \frac{(n_b - n_a + n_{a\bar{b}})^2 + (n_{\bar{b}} - n_{a\bar{b}})^2}{n n_{\bar{a}}} - \frac{n_b^2}{n^2} - \frac{n_{\bar{b}}^2}{n^2}$	[30]
15	II	$1 - \sum_{k=max(0, n_a - n_b)}^{n_{a\bar{b}}} \frac{C_{n_b}^{n_a - k} C_{n_{\bar{b}}}^{k}}{C_n^{n_a}}$	[13]
16	IPEE	$1 - \frac{1}{2^{n_a}} \sum_{k=0}^{n_{a\bar{b}}} C_{n_a}^{k}$	[7]
17	Jaccard	$\frac{n_a - n_{a\bar{b}}}{n_b + n_{a\bar{b}}}$	[30]
18	J-measure	$\frac{n_a - n_{a\bar{b}}}{n} log_2 \frac{n(n_a - n_{a\bar{b}})}{n_a n_b} + \frac{n_{a\bar{b}}}{n} log_2 \frac{n n_{a\bar{b}}}{n_a n_{\bar{b}}}$	[30]
19	Kappa	$\frac{2(n_a n_{\bar{b}} - n n_{a\bar{b}})}{n_a n_{\bar{b}} + n_{\bar{a}} n_b}$	[30]
20	Klosgen	$\sqrt{\frac{n_a - n_{a\bar{b}}}{n}}(\frac{n_{\bar{b}}}{n} - \frac{n_{a\bar{b}}}{n_a})$	[30]
21	Laplace	$\frac{n_a + 1 - n_{a\bar{b}}}{n_a + 2}$	[30]
22	Least Contradiction	$\frac{n_a - 2n_{a\bar{b}}}{n_b}$	[4]
23	Lift	$\frac{n(n_a - n_{a\bar{b}})}{n_a n_b}$	[26]
24	Lerman	$\frac{n_a - n_{a\bar{b}} - \frac{n_a n_b}{n}}{\sqrt{\frac{n_a n_b}{n}}}$	[13]
25	Loevinger	$1 - \frac{n n_{a\bar{b}}}{n_a n_{\bar{b}}}$	[22]
26	Odds Ratio	$\frac{(n_a - n_{a\bar{b}})(n_{\bar{b}} - n_{a\bar{b}})}{n_{a\bar{b}}(n_b - n_a + n_{a\bar{b}})}$	[30]
27	Pavillon/Added Value	$\frac{n_{\bar{b}}}{n} - \frac{n_{a\bar{b}}}{n_a}$	[30]
28	Phi-Coefficient	$\frac{n_a n_{\bar{b}} - n n_{a\bar{b}}}{\sqrt{n_a n_b n_{\bar{a}} n_{\bar{b}}}}$	[30]
29	Putative Causal Dependency	$\frac{3}{2} + \frac{4n_a - 3n_b}{2n} - (\frac{3}{2n_a} + \frac{2}{n_{\bar{b}}})n_{a\bar{b}}$	[18]
30	Rule Interest	$\frac{n_a n_{\bar{b}}}{n} - n_{a\bar{b}}$	[25]
31	Sebag & Schoenauer	$\frac{n_a}{n_{a\bar{b}}} - 1$	[28]
32	Support	$\frac{n_a - n_{a\bar{b}}}{n}$	[1]
33	TIC	$\sqrt{DIR(a \rightarrow b) \times DIR(\bar{b} \rightarrow a)}$	[8] [6]
34	Yule's Q	$\frac{n_a n_{\bar{b}} - n n_{a\bar{b}}}{n_a n_{\bar{b}} + (n_b - n_{\bar{b}} - 2n_a)n_{a\bar{b}} + 2n_{a\bar{b}}^2}$	[30]
35	Yule's Y	$\frac{\sqrt{(n_a - n_{a\bar{b}})(n_{\bar{b}} - n_{a\bar{b}})} - \sqrt{n_{a\bar{b}}(n_b - n_a + n_{a\bar{b}})}}{\sqrt{(n_a - n_{a\bar{b}})(n_{\bar{b}} - n_{a\bar{b}})} + \sqrt{n_{a\bar{b}}(n_b - n_a + n_{a\bar{b}})}}$	[30]

Annexe B

Interactive visual exploration of association rules with the rule focusing methodology [22]

In journal : *Knowledge and Information Systems*, Springer, 2007.

Under consideration for publication in Knowledge and Information Systems

Interactive Visual Exploration of Association Rules with Rule Focusing Methodology

Julien Blanchard, Fabrice Guillet and Henri Briand

KnOwledge & Decision Team (KOD)
LINA – FRE CNRS 2729, Polytechnic School of Nantes University, France

Abstract. On account of the enormous amounts of rules that can be produced by data mining algorithms, knowledge post-processing is a difficult stage in an association rule discovery process. In order to find relevant knowledge for decision-making, the user (a decision-maker specialized in the data studied) needs to rummage through the rules. To assist him/her in this task, we here propose the *rule focusing* methodology, an interactive methodology for the visual post-processing of association rules. It allows the user to explore large sets of rules freely by focusing his/her attention on limited subsets. This new approach relies on rule interestingness measures, on a visual representation, and on interactive navigation among the rules. We have implemented the rule focusing methodology in a prototype system called *ARVis*. It exploits the user's focus to guide the generation of the rules by means of a specific constraint-based rule-mining algorithm.

Keywords: knowledge discovery in databases, association rules, post-processing, interactive visualization, rule focusing, constraint-based mining, interestingness measures, neighborhood of rules

1. Introduction

Among the knowledge models used in Knowledge Discovery in Databases (KDD), association rules (Agrawal et al, 1993) have become a major concept and have received significant research attention. Association rules are implicative tendencies $X \rightarrow Y$ where X and Y are conjunctions of items (boolean variables of the form *databaseAttribute=value*). The left-hand side X is the antecedent of the rule and

Received Oct 7, 2004
Revised Nov 7, 2005
Accepted May 3, 2006

the right-hand side Y the consequent. Such a rule means that most of the records which verify the antecedent in the database verify the consequent too. For instance, in market basket analysis where the data studied are the customers' transactions in a supermarket, an association rule $\{pizza, crisps\} \rightarrow \{beer\}$ means that if a customer buys a pizza and crisps then (s)he most probably buys beer too. Since the pioneering algorithm of Agrawal, called Apriori (Agrawal and Srikant, 1994), many algorithms have been proposed for association rule mining (cf. Hipp et al (2000) for a survey). They generally produce very large amounts of rules. This is due to the unsupervised nature of association rule discovery. Indeed, because the user does not know precisely enough what (s)he is looking for to express it with the data terminology, (s)he does not make his/her goals explicit and does not specify any endogenous variable. Thus, the algorithms search all the valid associations existing in the database and generate an amount of rules exponentially growing with the number of items.

A crucial step in association rule discovery is post-processing, i.e., the interpretation, evaluation and validation of the rules in order to find interesting knowledge for decision-making. Because of the oversized amounts of rules, the post-processing stage often turns out to be a second mining challenge called "knowledge mining". While data mining is automatically computed by combinatorial algorithms, the knowledge mining stage is manually done by the user (a decision-maker specialized in the data studied). In practice, it is very difficult for users to rummage through the rules and find interesting ones in a corpus that can hold hundreds of thousands of rules, or even millions of rules with large business databases.

Many authors have stressed that the KDD process is by nature highly iterative and interactive and requires user involvement (Silberschatz and Tuzhilin, 1996) (Fayyad et al, 1996). In particular, Brachman and Anand (1996) have pointed out that in order to efficiently assist the users in their search for interesting knowledge, the KDD process should be considered not from the point of view of the discovery algorithms but from that of the users', as a human-centered decision support system. The human-centered approaches aim at creating a retroaction loop between the user and the system which constantly takes into account the information processing capacities of the user (cf. Bisdorff (2003) for examples of applications). Adopting Brachman & Anand's point of view, in this article we propose the *rule focusing* methodology, a human-centered methodology for the post-processing of association rules. The rule focusing methodology allows the user to explore large sets of rules by focusing his/her attention on successive limited subsets. The methodology relies on several *neighborhood relations* that connect the rules among them according to the user's semantics. With these relations, the user can navigate freely among the subsets of rules and thus drive the post-processing. In this way, a voluminous set of rules is explored subset by subset so that the user does not need to appropriate it entirely. Our approach combines:

- rule interestingness measures to filter and sort the rules,
- a visual representation to make comprehension easier,
- interactivity based on the neighborhood relations to guide the post-processing.

The rule focusing methodology can be used in two ways. First, it can be applied after association rule mining, as a pure post-processing technique. This is also called post-analysis or a posteriori filtering of rules (Hipp and Gntzer, 2002).

Secondly, it can be applied during association rule mining, as an interactive mining technique conducted by the user. Effectively, the rule focusing methodology induces a constraint-based rule-mining algorithm. A constraint-based rule-mining algorithm exploits constraints that the user gives to specify which kind of rules (s)he wants to find (cf. for example Srikant et al (1997), Ng et al (1998), Goethals and Van den Bussche (2000), Jeudy and Boulicaut (2002), Ordonez et al (2006)). Syntactic constraints (constraints specifying the items that must occur or not in the rule) and interestingness measure threshold constraints are the most commonly used constraints, but more general studies concern the so-called anti-monotone and succinct constraints (Ng et al, 1998), and the monotone constraints (Grahne et al, 2000) (Bonchi et al, 2005). Constraints allow to significantly reduce the exponentially growing search space of association rules[1]. Thus, the constraint-based algorithms can mine dense data more efficiently than the classical Apriori-like algorithms (the FP-growth-based algorithms of (Han et al, 2000) can mine dense data too, but they use a condensed representation of the data and require that it holds in memory). Besides, with appropriate constraints, the constraint-based algorithms can discover very specific rules which cannot be mined by the Apriori-like algorithms (the constraint-based algorithms can use low support thresholds for which Apriori-like algorithms are intractable). These rules are often very valuable for the users because they were not even thought of beforehand (Freitas, 1998). For these reasons, we use a constraint-based rule-mining algorithm to implement the rule focusing methodology in the prototype system described in this article. This specific algorithm extracts the rules interactively according to the user's focus. Note that using the rule focusing methodology as a pure post-processing technique or as an interactive mining technique is only a choice of implementation. The methodology does not depend on it.

The remainder of this article is organized as follows. In the next section we present a survey on association rule evaluation, exploration, and visualization. Then we describe the Information Visualization field of research, and in particular we compare 2D and 3D visualizations. Section 4 is dedicated to the study of cognitive constraints of the user during rule post-processing. From these constraints, in section 5, we define the rule focusing methodology. Section 6 describes the prototype system implementing our methodology: *ARVis*, a visual tool for association rule mining and post-processing . In section 7, we give an example of rule post-processing with *ARVis*. It comes from a study made with the firm PerformanSe SA on human resource management data. Finally we give our conclusion in section 8.

2. Survey on association rule evaluation, exploration, and visualization

At the output of the data mining algorithms, the sets of association rules are simple text lists. Each rule consists of a set of items for the antecedent, a set of items for the consequent (sets of items are called itemsets), and the numerical values of two interestingness measures, support and confidence (Agrawal et al, 1993). Support is the proportion of records which verify a rule in the database;

[1] Choosing the best way of harnessing multiple constraints whatever the data is still an open problem.

it evaluates the generality of the rule. Confidence (or conditional probability) is the proportion of records which verify the consequent among those which verify the antecedent; it evaluates the validity of the rule (success rate).

Three kinds of approaches aim at helping the user appropriate large sets of association rules:

- the user can filter and order the rules with other interestingness measures;
- the user can browse the large sets of rules with interactive tools or query languages;
- the user can visualize the rules.

2.1. Rule interestingness measures

It is now well-known that the support-confidence framework is rather poor to evaluate the rule quality (Silverstein et al, 1998) (Bayardo and Agrawal, 1999) (Tan et al, 2004). Numerous rule interestingness measures have been proposed to complement this framework. They are often classified into two categories: the subjective (user-oriented) ones and the objective (data-oriented) ones. Subjective measures take into account the user's a priori knowledge of the data domain (Liu et al, 2000) (Silberschatz and Tuzhilin, 1996) (Padmanabhan and Tuzhilin, 1999). On the other hand, the objective measures do not depend on the user but only on objective criteria such as data cardinalities or rule complexity. Depending on whether they are symmetric (invariable by permutation of antecedent and consequent) or not, they evaluate correlations or rules.

There exist two significant configurations in which the rules appear non-directed relations and therefore can be considered as neutral or non-existing (Blanchard, 2005):

- the *independence*, i.e., when the antecedent and consequent are independent;
- what we call the *equilibrium*, i.e., when examples and counter-examples are equal in numbers (maximum uncertainty of the consequent given that the antecedent is true).

Thus we distinguish two different but complementary aspects of the rule interestingness: the deviation from independence and the deviation from equilibrium. The objective measures of interestingness can be classified into two classes (Blanchard et al, 2005) (Blanchard et al, 2005):

- the measures of deviation from independence, which have a fixed value at independence, such as rule-interest (Piatetsky-Shapiro, 1991), lift (Silverstein et al, 1998), conviction (Brin et al, 1997), Loevinger index (Loevinger, 1947), implication intensity (Gras, 1996) (Blanchard et al, 2003);
- the measures of deviation from equilibrium, which have a fixed value at equilibrium, such as confidence (Agrawal et al, 1993), Sebag and Schoenauer index (Sebag and Schoenauer, 1998), IPEE (Blanchard et al, 2005).

These two kinds of measures are complementary (in particular, they do not create the same preorder on rules) (Blanchard, 2005) (Blanchard et al, 2005). A rule can have a good deviation from independence with a bad deviation from equilibrium, and conversely. Regarding the deviation from independence, a rule $A \rightarrow B$ with a good deviation means "When A is true, then B is **more** often true" (more than usual, i.e. more than without any information about A). On

the other hand, regarding the deviation from equilibrium, a rule $A \rightarrow B$ with a good deviation means "When A is true, then B is **very** often true". Deviation from independence is a comparison relatively to an expected situation, whereas deviation from equilibrium is an absolute statement. The measures of deviation from independence are useful to discover associations between antecedent and consequent (do the truth of A influence the truth of B?), while the measures of deviation from equilibrium are useful to take decisions or make predictions about B (knowing or supposing that A is true, is B true or false?) (Blanchard, 2005).

2.2. Interactive browsing of rules

Interactive tools of the type "rule browser" have been developed to assist the user in the post-processing of association rules. First, Klemettinen et al (2004) present a browser with which the user reaches interesting rules by adjusting thresholds on interestingness measures and applying syntactic constraints (templates). Secondly, Liu et al (1999) propose a rule browser which is based on subjective interestingness measures and exploits the user's a priori knowledge of the data domain to present the rules. The user expresses his/her knowledge under the form of relations and then the tool classifies the rules in different categories according to whether they confirm or not the user's beliefs. Finally, in (Ma et al, 2000), the user explores a summary of the rules. (S)he can access the rules by selecting elements in the summary. The main limit of all these tools lies in the textual representation of the rules which does not suit the study of large amounts of rules described by numerous interestingness measures.

More recently, a rule browser equipped with numerous functionalities has been presented in (Fule and Roddick, 2004). It allows to filter the rules with syntactic constraints that are more or less general since they can take into account an item taxonomy. The tool also enables the user to program any interestingness measure to order and filter the rules. Besides, the user can save the rules that (s)he judges interesting during the exploration. Another rule browser is presented in (Tuzhilin and Adomavicius, 2002), but it is not a generic tool. It is dedicated to the analysis of gene expression data coming from DNA microarrays, and relies on a very complete system of syntactic constraints which can take into account a gene taxonomy.

Within the framework of inductive databases (Imielinski and Mannila, 1996), several rule query languages have been proposed, such as DMQL (Han et al, 1996), MINE RULE (Meo et al, 1998), MSQL (Imielinski and Virmani, 1999), or XMINE (Braga et al, 2002). They allow to mine (by means of constraint-based algorithms) and post-process rules interactively under the user's guidance. However, as regards rule post-processing, the query languages are not very user-friendly (cf. (Botta et al, 2002) for an experimental study).

2.3. Visualizing the rules

Visualization can be very beneficial to KDD (Fayyad et al, 2001). Visualization techniques are indeed an effective means of introducing human subjectivity into each stage of the KDD process while taking advantage of the human perceptual capabilities. The information visualization techniques can either be used as knowledge discovery methods on their own, which is sometimes called "visual

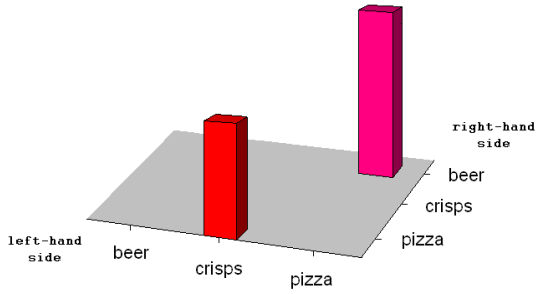


Fig. 1. An item-to-item matrix showing the rules $\{crisps\} \rightarrow \{pizza\}$ and $\{pizza\} \rightarrow \{beer\}$

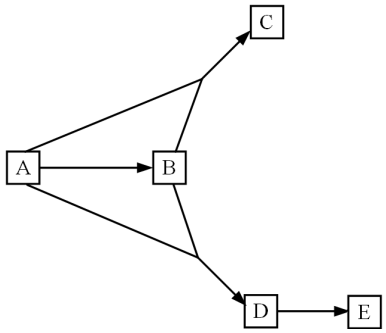


Fig. 2. A rule graph with items as nodes, showing the rules $A \rightarrow B$, $AB \rightarrow C$, $AB \rightarrow D$, and $D \rightarrow E$

data mining" (Keim, 2002), or they can collaborate with data mining algorithms to facilitate and speed up the analysis of data, intermediate results, or discovered knowledge (Aggarwal, 2002) (Schneiderman, 2002) (Han et al, 2003). Association rule visualization comes within this latter case. It must be noticed that the methods and tools presented below are generally supplied with basic functionalities for ordering and filtering the rules on items and on a few interestingness measures.

A first rule visualization method consists in using a matrix representation. Hofmann and Wilhelm (2001) and the Quest research group (Agrawal et al, 1996), as well as the software programs DBMiner (Han et al, 1997), SGI MineSet (Brunk et al, 1997), DB2 Intelligent Miner Visualization (**?**), and Enterprise Miner (**?**), give different implementations of it. In an item-to-item matrix (figure 1), each line corresponds to an antecedent item and each column to a consequent item. A rule between two items is symbolized in the matching cell by a 2D or 3D object whose graphical characteristics (generally size and color) represent the interestingness measures. This visualization technique has been improved into rule-to-item matrices (Wong et al, 1999) whose cluttering is lower and which allow a more efficient representation of rules with more than two items. The main limit of these approaches is that the matrices reach considerable sizes in case of large sets of rules over numerous items.

Sets of association rules can be also visualized by using a directed graph (Klemettinen et al, 2004) (Han et al, 1997) (Rainsford and Roddick, 2000) (**?**), the nodes and edges respectively representing the items and the rules (cf. figure

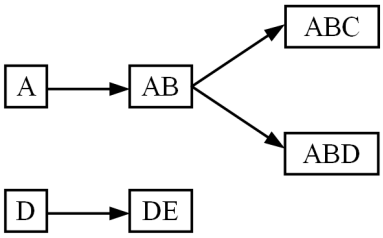
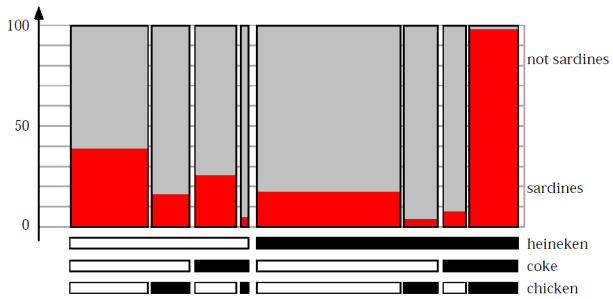


Fig. 3. A rule graph with itemsets as nodes, showing the rules $A \rightarrow B$, $AB \rightarrow C$, $AB \rightarrow D$, and $D \rightarrow E$



The represented rules involve the items *heineken*, *coke*, and *chicken* in antecedent, and *sardines* in consequent. Antecedent and consequent are respectively read on the horizontal and vertical axes. The first rectangle on the left shows that the rule $\{heineken = 0, coke = 0, chicken = 0\} \rightarrow \{sardines = 1\}$ has a confidence of about 40% (red part of the rectangle), while the opposite rule $\{heineken = 0, coke = 0, chicken = 0\} \rightarrow \{sardines = 0\}$ has a confidence of 60% (grey part of the rectangle). The support of the rules is proportional to the area (red or grey) of the rectangles. The rectangle on the right indicates that the confidence of the rule $\{heineken = 1, coke = 1, chicken = 1\} \rightarrow \{sardines = 1\}$ is about 100%.

Fig. 4. Mosaic display for association rules (from Hofmann et al (2000))

2 where letters denote items). The interestingness measures are symbolized on the edges, for instance with color or thickness. The graph representation is very intuitive but it has two main drawbacks. First, it makes transitivity appear among the rules whereas, in the general case, the rules are not transitive (with most measures, rule interestingness does not spread transitively). Secondly, it does not suit the visualization of large sets of rules over numerous items either. Indeed, the graph is then overloaded with nodes and crossing edges, all the more when rules with more than two items are considered. To improve the rule visualization, the same representation method has been used in 3D with a self-organization algorithm to guarantee a more efficient graph layout (Hao et al, 2001). Also we have proposed in (Kuntz et al, 2000) a dynamic rule graph which is a subgraph of the itemset lattice. In this graph, the nodes do not represent the items but the itemsets so that a rule $AB \rightarrow C$ is symbolized by an edge between the nodes AB and ABC (figure 3). The resulting graph is acyclic with more nodes but fewer edge crossings. The user can dynamically develop the graph as (s)he wishes by clicking on the nodes.

All the visual representations described so far are based on rule syntax (i.e. the items). A different approach is proposed in (Unwin et al, 2001), where the

representation is based on interestingness measures. This representation is a scatterplot between support and confidence where each point is colored according to density estimation. The user can query any point to display the names of the rules represented by the point (rules with close supports and confidence). The main advantage of such a representation is that it can contain a great number of rules. However, several rules can be represented by one and only one point, which does not facilitate the task of the users when they search for rules using items as criteria. This approach is the closest to the one we propose in this paper, which also uses a spatial mapping to highlight the interestingness measures.

Other methods have been proposed to represent association rules. Nevertheless, they do not deal with the visualization of the whole rule set but with the visualization of a pattern of rules (a group of rules with given items in antecedent and consequent). These methods allow a thorough study of a restricted number of rules, making their interpretation easier and helping to understand their occurrence context. We can quote for example Hofmann et al.'s mosaic plots (2000) for rules with categorical attributes (figure 4), or Fukuda et al (2001) and Han et al (2003) for numerical rules. Also some techniques inspired from parallel coordinates have been considered to visualize patterns of classification rules (Han et al, 2000) or association rules (Kopanakis and Theodoulidis, 2001).

3. Information visualization

3.1. Context

Information visualization (Card et al., 1999) (Spence, 2000) consists in representing abstract data under a visual form in order to improve cognition for a given task, that is to say the acquisition and use of new knowledge. The core of information visualization is visual encoding, i.e., the mapping of data tables to visual structures in a 2D or 3D space (Card et al., 1999). The visual structures have several graphical properties such as position, length, area, hue, brightness, saturation, shape, texture, angle, curvature... They can be zero-dimensional (points), one-dimensional (lines), two-dimensional (surfaces), or three-dimensional (volumes).

Several authors proposed classifications of visual encodings in order to show which ones are appropriate according to the data variables to be represented. Among these works, those of Cleveland and McGill (1984), Tufte (1983), and then Wilkinson (1999) are references for statistical graphs (charts). A second trend stems from cartography, with the works of McEachren (1995) and Bertin whose *Semiology of Graphics* (Bertin, 1967-1983) is considered as the first and most influential structural theory of graphics (Wilkinson, 1999). As regards the visual representation of quantitative variables, the two trends agree that the best encodings are done with position (Bertin, 1967-1983) (Cleveland and McGill, 1984) (McEachren, 1995) (Card et al., 1999) (Wilkinson, 1999). However, the two trends mainly differ about the use of surfaces to represent quantitative variables: this use is not advisable with statistical graphs whereas it is standard practice in cartography. In particular, Cleveland and McGill (1984) propose a hierarchy of visual encodings saying that surface is little appropriate (less than length) to represent quantities. This point of view is based on Stevens's law in psychophysics according to which the perceived quantities are not linearly related to the actual quantities with surface (Baird, 1970). On the other hand, Bertin points out that

before the variation of length, the variation of surface is the sensitive stimulus of the variation of size (Bertin, 1967-1983).

The visualization we propose in this article is not a map, and even less a statistical graph: this is a 3D virtual world. With the increase in the capacities of personal computers, the 3D virtual worlds have become common in information visualization (Chen, 2004). Associated with navigation operators (viewpoint controls), they have shown to be efficient for browsing wide information corpuses such as large file system hierarchies with Silicon Graphics' FSN (re-used in MineSet for the visualization of decision trees), hypertext document graphs with Harmony (Andrews, 1995), or OLAP cubes with DIVE-ON (Ammoura et al, 2001)] (cf. Chen (2004) for other examples of applications). While a 2D representation is restricted to the two dimensions of the screen, the additional dimension in a 3D virtual world offers a viewpoint towards infinity, creating a wide workspace capable of containing a large amount of information (Card et al., 1999). In this workspace, the most important information can be placed in the foreground (most visible objects) and thus be highlighted compared to the less important information placed behind it (less visible objects). This is the reason why the 3D representations are sometimes considered as focus+context approaches. Moreover, 3D enables to exploit volumes as objects in the visualization space. It allows to benefit from more graphical properties for the objects and thus to represent even more information.

3.2. 2D or 3D?

The choice between 2D and 3D representations for information visualization is still an open problem (Card et al., 1999) (Chen, 2004). This is especially due to the fact that the efficiency of a visualization is highly task-dependent (Carswell et al, 1991). Besides, while 3D representations are often more attractive, 2D has the advantage of a long and fruitful experience in information visualization. In fact, few research works are dedicated to the comparison between 2D and 3D. As regards the static (non interactive) visualization of statistical graphs, the 3D representations have generally not been advisable since the influential publications of Tufte (1983) and Cleveland and McGill (1984). Nevertheless, the psychophysics experiments of Spence (1990) and Carswell et al (1991) show that there is no significant difference of accuracy between 2D and 3D for the comparison of numerical values. In particular, Spence points out that this is not the apparent dimensionality of visual structures which counts (2 for a surface, 3 for a volume) but the actual number of parameters that show variability (Spence, 1990). In his experiments, whatever the apparent dimensionality of visual structures, Stevens's law is almost always the same when only one parameter actually varies (Stevens's law exponents are very close to 1). Under some circumstances, information may even be processed faster when represented in 3D rather than in 2D. As regards the perception of global trends in data (increase or decrease), the experimental results of Carswell et al (1991) also show an improvement in the answer times with 3D but to the detriment of accuracy.

Other works compare 2D and 3D within the framework of interactive visualization. Cockburn and McKenzie (2001) study the storage and retrieval of bookmarked web-pages in a 2D or 3D visualization space. With the 2D interface, the processing times of the users are shorter but not significantly. On the other hand, the subjective assessment of the interfaces shows a significant preference

for 3D (which Spence (1990) and Carswell et al (1991) also sense but without assessing it). Finally, Ware and Franck (1996) compare the visualization of 2D graphs and 3D graphs. Their works show a significant improvement in intelligibility with 3D. More precisely, their experiment consists in asking users whether there is a path of length two between two nodes randomly chosen in a graph. With the 3D graphs, the error rate is reduced by 2.2 for comparable answer times. With stereoscopic display, the error rate is even reduced by 3. One generally considers that only stereoscopy allows fully exploiting the characteristics of the 3D representations.

4. Cognitive constraints of the user during rule post-processing

4.1. User's task

During the post-processing of the rules, the user is faced with long lists of rules described by interestingness measures. The user's task is then to rummage through the rules in order to find interesting ones for decision-making. To do so, (s)he needs to interpret the rules in the business semantics and to evaluate their quality. The two decision indicators are therefore the rule syntax and the interestingness measures. The user's task is difficult for two reasons. First, the profusion of rules at the output of the data mining algorithms prevents any exhaustive exploration. Secondly, on account of the unsupervised nature of association rule discovery, it is generally not feasible for the user to obviously formulate constraints which would isolate relevant rules directly.

4.2. Cognitive hypotheses of information processing

On account of the human "bounded rationality" hypothesis (Simon, 1979), a decision process can be seen as a search for a dominance structure. More precisely, the decision-maker faced with a set of multiattribute alternatives tries to find an alternative (s)he considers dominant over the others, i.e., an alternative (s)he thinks better than the others according to his/her current representation of the decision situation (Montgomery, 1983). This type of models of decision process can be transferred to the post-processing of association rules by considering the rules as a particular kind of alternatives with items and interestingness measures as attributes. According to Montgomery, the decision-maker isolates a limited subset of potentially useful alternatives and makes comparisons among them. This can be done iteratively during the decision process. More precisely, he has pointed out that: "The decision process acquires a certain directionality in the sense that certain alternatives and attributes will receive more attention than others [...] The directionality of the process may be determined more or less consciously. Shifts in the directionality may occur several times in the process, particularly when the decision-maker fails to find a dominance structure".

Furthermore, a KDD methodology called "attribute focusing" has been proposed in (Bhandari, 1994). It results from experimental data concerning the user's behavior in the discovery process. This methodology is based on a filter which automatically detects a small number of potentially interesting attributes.

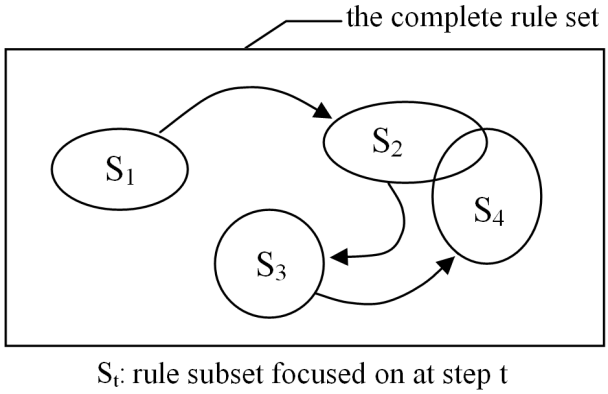


Fig. 5. Navigation among successive subsets of rules with the rule focusing methodology

The filter guides the user's attention on a small, and therefore more intelligible, subset of the database. The importance of focusing on a small number of attributes in human information processing has also been widely confirmed with works on decision strategies (cf. for example the moving basis heuristics in (Barthelemy and Mullet, 1992)). Indeed, on account of his/her limited cognitive abilities, the decision-maker examines only a small amount of information at each moment.

From these different works on human information processing, we establish three principles on which our rule focusing methodology relies:

P1. enabling the user to focus his/her attention on a limited subset of rules with a small number of attributes (items and interestingness measures),

P2. enabling the user to make comparisons among the rules in the subset,

P3. enabling the user to shift the subset of rules (s)he is focusing on at any time during the post-processing, until (s)he is able to validate some rules and reach a decision.

5. Rule focusing methodology

The idea of developing the *rule focusing* methodology has arisen from our earlier works on the visualization of rule sets by graphs (Kuntz et al, 2000). The methodology consists in letting the user navigate freely inside the large set of rules by focusing on successive limited subsets via a visual representation of the rules and their measures. In other words, the user gradually drives a series of visual local explorations according to his/her interest for the rules (figure 5). Thus, the rule set is explored subset by subset so that the user does not need to appropriate it entirely. At each navigation[2] step, the user must make a decision to choose which subset to visit next. This is the way subjectivity is introduced into the post-processing of the rules. The user acts here as an exploration heuristics.

[2] We call "navigation" the fact of going from one subset to another, while "exploration" refers to the whole process supervised by our methodology, i.e., the navigation among the subsets and the visits (local explorations) of the subsets.

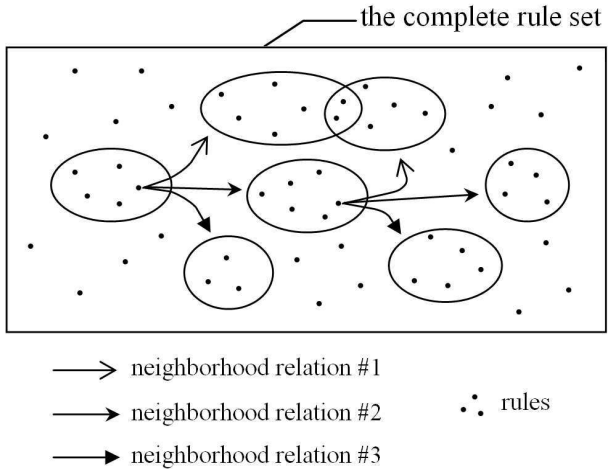


Fig. 6. A neighborhood relation associates each rule to a subset of rules

Exploiting a human heuristics is coherent since the function to be optimized, i.e. the user's interest, is subjective.

The rule focusing methodology integrates the cognitive principles of section 4.2 in the following way:

- Relations allow to focus on the subsets and to navigate among them (principles P1 and P3). We call them *neighborhood relations*.
- The user visualizes the subsets to visit them, and in particular to compare the rules (principle P2).

Both the neighborhood relations and the visualization technique must take into account the two decision indicators involved in the user's task: the rule syntax and the interestingness measures (cf. section 4.1).

5.1. Neighborhood relations among rules

The neighborhood relations determine the way the subsets of rules are focused on (cognitive principle P1) and the way the user can go from one subset to another (cognitive principle P3). They are a fundamental element of the rule focusing methodology since they are the vectors of the navigation for the user. These relations are defined in the following way: a neighborhood relation associates each rule from the complete set of rules to a limited subset of rules called neighbors (figure 6). So with x relations, the user can reach x neighboring subsets of rules from one rule, and from a subset containing y rules (s)he can reach $x.y$ possible neighboring subsets. To navigate from one subset to another, the user must make two choices: which neighborhood relation to apply, and on which rule.

In mathematical terms, the neighborhood relations are binary relations over the complete set R of the rules extracted by the data mining algorithms. Still with the aim of being appropriate to the user's task, we choose neighborhood relations which have a pertinent meaning for the user:

$\forall (r_1, r_2) \in R^2$,
$neighborOf(r_1, r_2) \Leftrightarrow$ (the user judges that r_1 is close to r_2 from a point of view)

This introduces user semantics into the navigation among the rules. Any relation could be considered provided it makes sense for the user. Consequently, the relations have to be defined with his/her help before starting the rule post-processing.

Here are for example four possible neighborhood relations $neighborOf(r_1, r_2)$:

1. r_1 is neighbor of r_2 if and only if r_1 and r_2 have the same conclusion;
2. r_1 is neighbor of r_2 if and only if r_1 is an exception of r_2;
3. r_1 is neighbor of r_2 if and only if the antecedent of r_1 is more general than that of r_2;
4. r_1 is neighbor of r_2 if and only if r_1 has the same support and confidence as r_2 to within about 0.05.

The neighborhood relations 1, 2, and 3 are based on the rule syntax, while relation 4 is based on two interestingness measures. Furthermore, relation 1 is an equivalence relation, whereas relation 2 is neither reflexive, nor symmetric, nor transitive. Relation 3 is only transitive, and relation 4 is reflexive and symmetric but not transitive.

Let us assume that the user applies a neighborhood relation Π. From a rule r, (s)he can reach the subset S of all the rules that are neighbors of r according to Π. We call r the "transitional rule" because it allows to navigate from one subset to another. Depending on the reflexivity of the relation Π chosen, S can or cannot contain the transitional rule r.

The originality of our methodology in comparison with the existing rule exploration techniques (described section 2.2) mainly lies in the concept of neighborhood relation. With a query language or an interactive interface like a rule browser, the user can reach any subset of rules but (s)he must explicitly specify the constraints which delimit it. With the rule focusing methodology, the constraint specification is implicit since it is hidden in the neighborhood relations. Actually, the neighborhood relations can be seen as generalizations of constraints (classes of constraints). We think that the user's task is made easier with neighborhood relations than with explicit constraints. For example, let us imagine the following exploration scenario:
The user finds an interesting rule $ABC \rightarrow D$ (where letters denote items). (S)he thinks that the combination of these four items is very pertinent but (s)he would like to change the order of the items and verify whether the rules $ABD \rightarrow C$, $ACD \rightarrow B$, and $BCD \rightarrow A$ (and why not the rules $AB \rightarrow CD$, $AC \rightarrow BD$ and so on as well) are better evaluated by the interestingness measures.

With the rule focusing methodology, the user can carry out this scenario in just one interaction. (S)he only needs the neighborhood relation "r_1 is neighbor of r_2 if and only if r_1 and r_2 have the same items". On the other hand, with a query language or a rule browser, the user has to write a series of appropriate queries or to specify a series of constraints manually with the graphical interface. This can be a tedious and time-consuming task.

5.2. Rule visualization

In order to help the user to visit the subsets of rules, we provide him/her with a visual representation instead of poorly intelligible textual lists. The visual representation facilitates and speeds up comprehension, and in particular it makes the comparisons among the rules easier (cognitive principle P2). Most of the techniques proposed for rule visualization have been developed to represent the whole set of rules produced by the data mining algorithms. Nevertheless, in the rule focusing methodology, we can take advantage of the user's focus strategy by representing only the current subset of rules at each navigation step. This reduces the number of rules to draw and above all largely improves the representation intelligibility. Visually, the user's point of view on the complete rule set is thus always local.

The rule visual representations are generally based on the rule syntax and handle interestingness measures as additional information (except for the approach of Unwin et al (2001), cf. section 2.3). However, the interestingness measures are also decision indicators fundamental to the user's task. So that the user can quickly assess and compare the rules, the representation must highlight the interestingness measures and make the best rules clearly recognizable. Also the visualization must be able to integrate numerous measures (not only support and confidence as it often happens), to dynamically filter the rules according to thresholds set by the user, and to support large amounts of rules having any number of items inside. Finally, the visualization must integrate interactive operators allowing the user to trigger the neighborhood relations.

Research works in visual perception show that a human being has first a global perception of a scene, before taking an interest in details. This is what motivated the development of the approaches named overview+details and focus+context (Card et al., 1999). Thus, in the rule focusing methodology, the user has to be able to easily change between global and detailed views of the rules by interacting with the visualization.

6. *ARVis*, a visual tool for association rule mining and post-processing

In this section, we present *ARVis (Association Rule Visualization)*, an experimental prototype implementing the rule focusing methodology. It was originally developed for the firm PerformanSe SA in order to find knowledge for decision support in human resource management. *ARVis* considers rules with single consequents (one item only in the consequent). This choice is usual in association rule discovery. Indeed, in association rule discovery in general and in our applications with PerformanSe SA in particular, the users are often interested spontaneously in this kind of rules because they are more intelligible than rules with multi-item consequents. However, considering only rules with single consequents is not a limitation to our approach. This choice could be easily changed.

At least three interestingness measures are calculated in *ARVis*: support, confidence (Agrawal et al, 1993), and entropic implication intensity (respectively noted sp, cf and eii). We choose support and confidence because they are the basic indexes to assess association rules. As for implication intensity, it is an asymmetric probabilistic index which evaluates the statistical significance

of the rules by quantifying the unlikelihood of the number of counter-examples (Guillaume et al, 1998) (Gras, 1996). The entropic version of this index also takes into account the imbalances between examples and counter-examples for both the rule and its contrapositive (Blanchard et al, 2003). The entropic implication intensity is a powerful measure since it takes into account both the deviation from independence and the deviation from equilibrium. This is the reason why we have chosen to integrate it into *ARVis*. But here again, this choice of measures is not a limitation to our approach, and others can be added. Each measure is associated to minimum and maximum thresholds set by the user: min_{sp}, min_{cf}, min_{eii}, max_{sp}, max_{cf}, max_{eii}. Although most of the tools for association rule mining do not provide them, the maximum thresholds improve the user's focus. For example, rules with high support and high confidence are often already known by the users; removing them allows highlighting more interesting rules.

In *ARVis*, we have opted for neighborhood relations mainly based on items and for a visualization technique mainly based on interestingness measures. We think this is the way to the most user-friendly solutions for rule exploration.

6.1. Neighborhood relations

Eight neighborhood relations are implemented in *ARVis*, most of them being generalization-type relations or specialization-type relations. Two of the most fundamental human cognitive mechanisms for generating new rules are indeed generalization and specialization (cf. the study of the reasoning processes in (Holland et al, 1986)).

Given the set I of items relative to the data studied, the rules are of the form $X \rightarrow y$ where X is an itemset $X \subset I$ and y is an item $y \in I - X$. The complete set of rules with single consequents that can be built with the items of I is noted R. In order to simplify the notations, we note $X \cup y$ instead of $X \cup \{y\}$ and $X - y$ instead of $X - \{y\}$. For the same simplicity reason, we define the neighborhood relations not as binary relations over R but as functions Π from R to 2^R which associate each rule with the subset composed of its neighbors:

$$\forall r_1 \in R, \Pi(r_1) = \{r_2 \in R \mid neighborOf(r_1, r_2)\}$$

Each of the eight neighborhood relations below induces two kinds of constraints:

- syntactic constraints, which specify the items that must occur or not in the antecedent and in the consequent;
- interestingness measure constraints, which specify minimum and maximum thresholds for the measures.

The syntactic constraints are peculiar to each neighborhood relation. On the other hand, the interestingness measure constraints are shared by all the relations. We group them together into the boolean function $interesting()$:

$$\forall r \in R, interesting(r) \Leftrightarrow \left| \begin{array}{l} min_{sp} \leq sp(r) \leq max_{sp} \\ min_{cf} \leq cf(r) \leq max_{cf} \\ min_{eii} \leq eii(r) \leq max_{eii} \end{array} \right.$$

A rule is said interesting if the three measures respect the minimum and maximum thresholds.

Specialization-type relations

$$agreement\_specialization(X \rightarrow y) = \left\{ X \cup z \rightarrow y \;\middle|\; \begin{array}{l} z \in I - (X \cup y) \\ interesting(X \cup z \rightarrow y) = true \end{array} \right\}$$

$$exception\_specialization(X \rightarrow y) = \left\{ X \cup z \rightarrow not(y) \;\middle|\; \begin{array}{l} z \in I - (X \cup\ not(y)) \\ interesting(X \cup z \rightarrow not(y)) \end{array} \right\}$$

$$forward\_chaining(X \rightarrow y) = \left\{ X \cup y \rightarrow z \;\middle|\; \begin{array}{l} z \in I - (X \cup\ y) \\ interesting(X \cup y \rightarrow z) \end{array} \right\}$$

Holland et al (1986) point out that a too general rule can be specialized into two kinds of complementary rules: exception rules and agreement rules. Exception rules aim at explaining the counter-examples of the general rule, while agreement rules aim at better explaining the examples. For instance, a rule "If α is a dog then α is friendly" can be specialized into the rules "If α is a dog and α is muzzled then α is mean" and "If α is a dog and α is not muzzled then α is friendly". The interest of exception rules in KDD has been widely confirmed (cf. for example (Hussain et al, 2000) (Suzuki, 2002)). On the basis of these two kinds of specialization, we propose the neighborhood relations *agreement_specialization* and *exception_specialization*[3] in *ARVis*. The third specialization-type relation is inspired by forward chaining in inference engines for expert systems: when a rule $X \rightarrow y$ is fired, the concept y becomes active and can be used with X to fire new rules and deduce new concepts z. Backward chaining cannot be considered with rules with single consequent.

Generalization-type relations

$$generalization(X \rightarrow y) = \left\{ X - z \rightarrow y \;\middle|\; \begin{array}{l} z \in X \\ interesting(X - z \rightarrow y) = true \end{array} \right\}$$

$$antecedent\_generalization(X \rightarrow y) = \left\{ X - z \rightarrow z \;\middle|\; \begin{array}{l} z \in X \\ interesting(X - z \rightarrow z) = true \end{array} \right\}$$

generalization relies on the condition-simplifying generalization mechanism described in (Holland et al, 1986). This relation is complementary to *agreement_specialization* and *exception_specialization*. It consists in deleting an item in the antecedent. The relation *antecedent_generalization* is complementary to *forward_chaining*. After applying *forward_chaining* on a rule r, one can effectively come back to r by applying *antecedent_generalization*.

Other relations

$$same\_antecedent(X \rightarrow y) = \left\{ X \rightarrow z \;\middle|\; \begin{array}{l} z \in I - X \\ interesting(X \rightarrow z) = true \end{array} \right\}$$

[3] To extend the notations to non-boolean attributes, $not(y)$ refers to any item coming from the same attribute as y but involving a different attribute value. For example, if y is the item *eye_color=blue* then $not(y)$ can be *eye_color=brown* or *eye_color=green*.

$$same\_consequent(X \rightarrow y) = \left\{ z \rightarrow y \;\middle|\; \begin{array}{l} z \in I - y \\ interesting(z \rightarrow y) = true \end{array} \right\}$$

$$same_items(X \rightarrow y) = \left\{ (X \cup y) - z \rightarrow z \;\middle|\; \begin{array}{l} z \in X \cup y \\ interesting((X \cup y) - z \rightarrow z) = true \end{array} \right\}$$

The relations *same_antecedent* and *same_consequent* preserve the antecedent and change the consequent, or vice versa. The relation *same_items* allows to reorder the items in a rule. All the rules produced by this relation concern the same population of records in the database.

6.2. Quality-oriented visualization

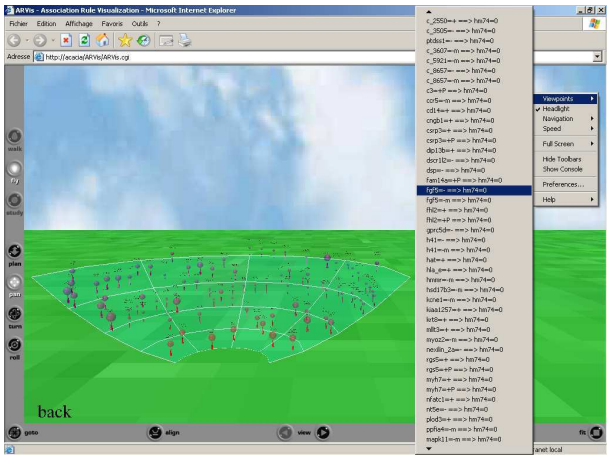
Each subset of rules is visualized in a 3D space, which we call a *world*. The representation is built in the following way: each rule is symbolized by an object composed of a sphere perched on top of a cone. Three straightforward graphical characteristics are thus obtained to represent the interestingness measures: the sphere diameter, the cone height, and the color. The representation size depends only on the number of rules in the subset and not on the amount of items. In order to facilitate the navigation (viewpoint control) inside the world, a ground and a sky are represented. As pointed out by Chen (2004), such visual landmarks make the navigation task easier by facilitating the acquisition of spatial knowledge, and more generally by facilitating the building of the cognitive map by the user (mental model of the world).

In a visual representation, the perceptually dominant information is the spatial position (Card et al., 1999). Therefore, in order to be emphasized, the interestingness measures which are fundamental for decision-making are represented by the object position in the world. Since several rules can present the same interestingness, only two measures can be symbolized by spatial position, so that the third dimension is free for scattering the objects. All things considered, we have chosen to use only one axis to place the objects in space and so to spatially represent only one interestingness measure. Indeed, the objects are laid out in the 3D world on an arena (a transparent half-bowl), which means that the further an object is, the higher it is placed (figure 7). This arena allows a better perception of the depth dimension and reduces occultation of objects by other objects. It can hold at most around 250 objects. A similar choice is made in the document manager Data Mountain of Microsoft Research, where web pages are laid out on an inclined plane (Robertson et al, 1998).

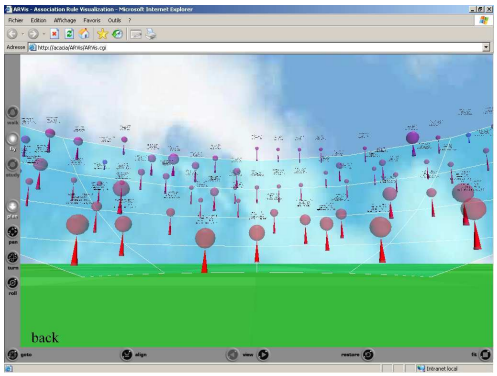
Weighing it all up, we have opted for the following visual metaphor to represent each subset of rules by highlighting the interestingness measures (figure 8):

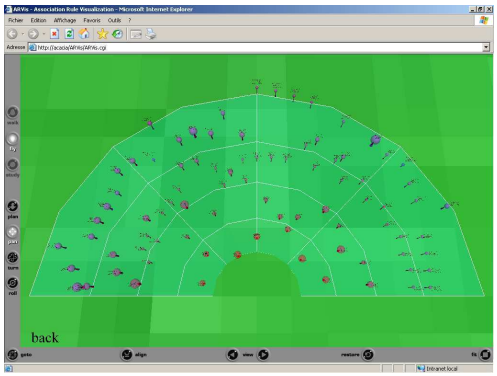
- the object position represents the entropic implication intensity,
- the sphere visible area represents the support,
- the cone height represents the confidence,
- the object color is used redundantly to represent a weighted average of confidence and entropic implication intensity, which gives a synthetic idea of the rule interestingness.

This visual metaphor facilitates comparisons among the rules. It stresses the good rules whose visualization and access are made easier compared to the less good



A





B C

Fig. 7. Each subset of rules is represented by a 3D world

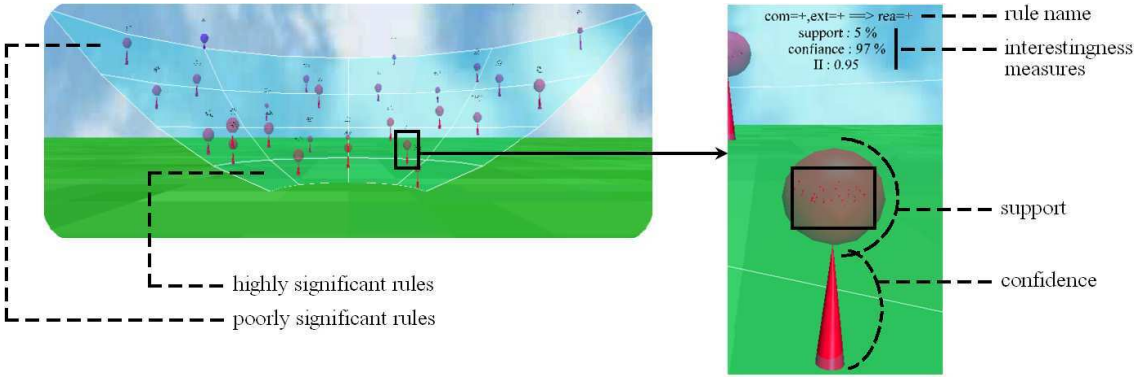
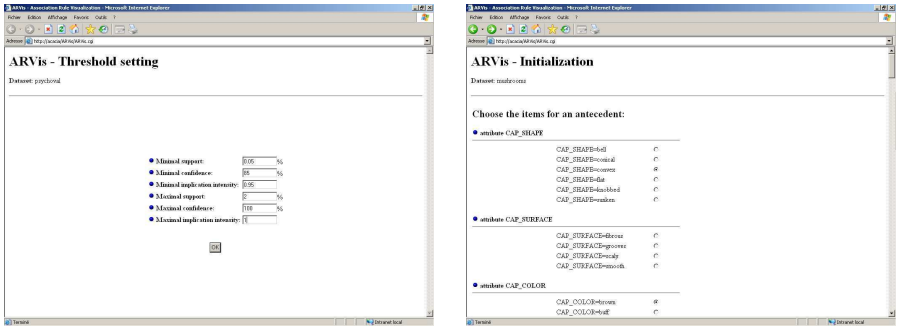


Fig. 8. Visual metaphor



A B

Fig. 9. Exploration initialization interface

rules. More precisely, a large red sphere perched on top of a tall cone placed at the front of an arena, on the lower steps, represents a rule whose support, confidence and entropic implication intensity are high. On the other hand, a little blue sphere perched on top of a small cone placed at the back, on the upper steps, represents a rule whose three measures are weak. This metaphor is a choice among the many possible combinations. It can be adapted by the user. One can choose for instance to change the mappings, or to represent more interestingness measures with color or by using two axes for the spatial mapping.

Furthermore, some complementary text labels appear above each object to give the name of the corresponding rule. They provide the numerical values for support, confidence and entropic implication intensity (noted "EII") too and thus complete the qualitative information given by the representation.

6.3. Interactions with the user

The user can interact in three different ways with the visual representation: by visiting a subset of rules, by filtering the rules on the interestingness measures in a subset, and by navigating among the subsets to discover new rules.

For each subset of rules, at the beginning of a visit, the user is placed in the 3D world in front of the arena so that (s)he benefits from an overall and synthetic view of the rules. With this comprehensive vision, it is easy to locate the best rules. Then the user can wander freely over the world to browse the rules, and zoom in on them to examine them more closely. (S)he just has to click on an object to move in front of it. In each 3D there also exist predefined viewpoints providing overall visions of the arena (cf. for instance, the viewpoint from the top in figure 7.C). If the user looks for a rule with particular items in it, (s)he can search it in a menu (figure 7.A) which lists all the rule names of the subset and allows to move directly in front of the object wanted. In the final analysis, *ARVis* enables the user to find the rules that interest him/her in a subset whether his/her search criteria are based on interestingness measures or on items.

At any time during the visit of a subset, the user can filter the rules on the interestingness measures by adjusting the thresholds min_{sp}, min_{cf}, min_{eii}, max_{sp}, max_{cf}, and max_{eii} (figure 9.A). Only the rules whose measures respect

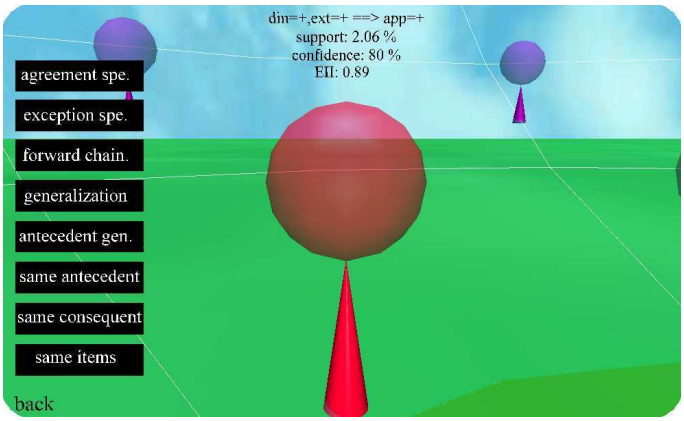


Fig. 10. For each rule, a menu can be displayed to choose the neighborhood relation to be triggered

the thresholds are represented. This makes objects appear or disappear in the world.

Finally, the user can navigate among the subsets of rules via a menu providing the eight neighborhood relations (figure 10). By applying a neighborhood relation on a rule, the current subset is replaced by a new subset. Visually, the current world is replaced by a new world, which gives the impression of virtually moving inside the whole set of rules. At any time during the navigation, the user can go back to the previous subsets (and worlds) with the "back" operator.

Let us assume that the user applies a neighborhood relation Π on a rule r. This generates a new subset $S = \Pi(r)$ containing all the rules neighbors of r according to Π. In *ARVis*, we systematically add the transitional rule r into the new subset S. Visually, the transitional rule r can be easily located in the world since its object flashes. This enables the user to make comparisons between the transitional rule r and its neighbor rules. For instance, with the neighborhood relation *agreement_specialization*, it is interesting to compare a rule r to its neighbors in order to see whether or not the addition of a new item in r tends to improve the rule interestingness. Reciprocally, with the relation *generalization*, comparing a rule r to its neighbors allows to detect the superfluous items in r (those whose removal does not reduce the quality of the rule).

To start or restart the navigation among the subsets, the user can choose the first subset to focus on with an exploration initialization interface (figure 9.B). This interface is an "itemset browser" working with inclusive templates: it enables to build the itemset of one's choosing and then to display the world of the rules that include this itemset in the antecedent, or in the consequent, or in both. Furthermore, the exploration initialization interface allows to choose the database and the table to be studied, and to choose the set I of the items to be used during the exploration.

6.4. *ARVis* implementation

6.4.1. Constraint-based rule-mining algorithm

When the user triggers a neighborhood relation, *ARVis* runs a constraint-based algorithm which dynamically computes the appropriate subset of rules with the interestingness measures. As seen in section 6.1, each of the eight neighborhood relations induces two kinds of constraints: syntactic constraints and interestingness measure constraints. These constraints are "pushed" into association rule mining to reduce the exponentially growing search space. The general structure of the constraint-based algorithm is given below.

(1) **Procedure** LocalMining
(2) Input: $rule$ //the transitional rule
(3) Π //the neighborhood relation
(4) $thresholds$ //the 6 thresholds on interestingness measures
(5) $database$ //connection to the database
(6) Output: $subset_{rules,measures}$ //subset of rules with interestingness measures

(7) $subset_{rules} = \emptyset$ //subset of rules (without measures)
(8) $cardinalities = \emptyset$ //cardinalities of the itemsets

(9) //STEP 1: Generate the candidate rules with the syntactic constraints
(10) $subset_{rules}$ = SyntacticGeneration($rule$,Π)

(11) //STEP 2: Count the itemsets of the candidate rules (database scan)
(12) $cardinalities$ = RetrieveCardinalities($subset_{rules}$,$database$)

(13) //STEP 3: Calculate the interestingness measures
(14) $subset_{rules,measures}$ = CalculateMeasures($subset_{rules}$,$cardinalities$)

(15) //STEP 4: Eliminate the candidate rules which do not respect
(16) //the interestingness measure constraints
(17) $subset_{rules,measures}$ = Filter($subset_{rules,measures}$,$thresholds$)
(18) **return**($subset_{rules,measures}$)

Only step 1 depends on the neighborhood relation Π chosen. This step needs no database scan since it deals only with the syntax of the rules. The syntactic constraints induced by the neighborhood relations of *ARVis* are powerful constraints which drastically reduce the number of rules to be produced. Effectively, the syntactic constraints are verified by at most $|I|$ rules, whatever the relation chosen. It is easy to enumerate these candidate rules and therefore to enumerate all the itemsets that must be counted in the database during step 2 (Ng et al (1998) pointed out the interest of such itemset enumeration procedures in constraint-based itemset-mining). Thus, whatever the neighborhood relation chosen, the whole constraint-based algorithm has a polynomial complexity[4] in $O(|I|)$.

In this polynomial algorithm, the most time-consuming step is step 2. It consists in counting the cardinalities of the itemsets by scanning the database. To improve the response times of the algorithm, a *progressive save system* is

[4] Except for the neighborhood relation *exception_specialization* for which the number of rules in a subset is bounded by $m.|I|$ and the complexity is polynomial in $O(m.|I|)$ where $m < |I|$ is the maximum of values for the attributes.

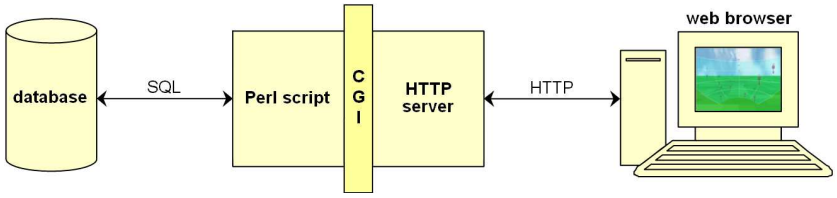


Fig. 11. *ARVis* architecture

implemented in the procedure RetrieveCardinalities: each time an itemset is counted, its cardinality is saved to avoid counting it another time during the remainder of the exploration. In this way, the greater number of times the algorithm is run over the same database, the more the itemset cardinalities are saved, and the more probable it is that the algorithm runs faster. The cardinalities of the itemsets are saved in the database in specific tables. There is one table for each itemset length (number of items in the itemset), so that all the itemsets of the same length are saved in the same table. For each table, the retrieval of the cardinalities uses a B-tree index.

Furthermore, for most of the neighborhood relations of *ARVis*, the constraint-based algorithm can be optimized by "pushing" the interestingness measure constraints into step 1. With the progressive save system, one can indeed anticipate that some candidate rules do not respect the thresholds. Eliminating these candidate rules allows to reduce the number of itemsets that must be counted in the database during step 2. For example, with the neighborhood relation *same_consequent*, it is useless to generate a candidate rule $antecedent \rightarrow consequent$ if the cardinality of *antecedent* has already been counted and does not respect $cardinality \geq n{*}min_{sp}/max_{cf}$ and $cardinality \leq n{*}max_{sp}/min_{cf}$ where n is the number of transactions in the database.

6.4.2. Architecture

ARVis is built on a client/server architecture with a thin client (figure 11). The main block is a CGI program in Perl divided into two parts:

- the constraint-based algorithm which dynamically extracts the subset of rules with their interestingness measures from the database,
- a procedure which dynamically generates the corresponding 3D world in VRML (this procedure is not time-consuming since no database access is needed).

The user visits the worlds with a web browser equipped with a VRML plug-in. The exploration initialization interface is a series of web pages generated by the CGI program.

6.4.3. Response time

The figure 12 shows the response times obtained on three datasets (presented table 1) by executing an exploration scenario with *ARVis*, i.e., a series of neighborhood relations. For each relation triggered by the user, the response time is the time required by *ARVis* to compute the subset of rules with the constraint-based algorithm and to display the corresponding world on the screen. The minimum and maximum thresholds chosen in the scenarios are given in table 2. For the

	# of items	**# of transactions**	**average transaction length**
MUSHROOMS	119	8416	23
T10.I4.D100k	100	100000	10
T20.I6.D100k	40	100000	20

Table 1. Dataset characteristics

	MUSHROOMS	**T10.I4.D100k**	**T20.I6.D100k**
min_{sp}	1%	0.05%	0.05%
max_{sp}	100%	100%	100%
min_{cf}	70%	70%	70%
max_{cf}	100%	100%	100%
min_{eii}	0.5	0	0
max_{eii}	1	1	1

Table 2. Minimum and maximum thresholds used in the exploration scenarios

experiments, the server of *ARVis* was running on an SGI Origin 2000 server equipped with four 250 MHz RISC R10000 processors and with 512 MB of memory. The DBMS was PostgreSQL. The tables storing the itemset cardinalities were empty at the beginning of the scenarios.

The first dataset is the MUSHROOMS dataset from the UCI Repository (Blake and Merz, 1998). It is small but it is known to be highly correlated. The exploration scenario that was used with this dataset is given in table 3. The two other datasets are two large synthetic ones: T10.I4.D100k and T20.I6.D100k. They were generated by the procedure proposed by Agrawal and Srikant (1994) (the number of patterns was set to 1000). The dataset T20.I6.D100k is deliberately very dense (on average, each transaction contains 43 % of the items). The exploration scenarios for these two datasets are similar to the one given in the table 3 but we do not present them since the data have no real meaning.

As can be seen on figure 12, the response times tend to decrease as the scenario unfolds. This is due to the progressive save system of the constraint-based algorithm of *ARVis*. The peaks in the response time curve (for example for t=6 and t=11 in the MUSHROOMS scenario) appear when the algorithm needs lots of itemsets that have not been counted yet. In this case, like in any classical procedure for frequent itemset mining, the algorithm has to scan the database, which is time-consuming.

The experiment on the dataset T20.I6.D100k shows that *ARVis* can efficiently mine dense data. In particular, during this experiment, very specific rules containing up to 15 items and presenting a support of 0.07 % have been computed. With a levelwise exhaustive algorithm, such specific rules could never be extracted from a dense database.

Time	Neighborhood relation	Transitional rule (on which the neighborhood relation is applied)	Number of rules generated
t=1	*forward_chaining*	$CLASS = edible \rightarrow GILL\_SIZE = broad$	3
t=2	*forward_chaining*	$CLASS = edible, GILL\_SIZE = broad \rightarrow ODOR = none$	4
t=3	*same_items*	$CLASS = edible, GILL\_SIZE = broad, ODOR = none \rightarrow STALK\_SHAPE = tapering$	3
t=4	*antecedent_generalization*	$GILL\_SIZE = broad, ODOR = none, STALK\_SHAPE = tapering \rightarrow CLASS = edible$	3
t=5	*same_antecedent*	$GILL\_SIZE = broad, STALK\_SHAPE = tapering \rightarrow ODOR = none$	7
t=6	*same_consequent*	$GILL\_SIZE = broad, STALK\_SHAPE = tapering \rightarrow RING\_NUMBER = one$	54
t=7	*forward_chaining*	$CLASS = edible, GILL\_SIZE = broad, STALK\_SHAPE = tapering \rightarrow ODOR = none$	10
t=8	*back* + *antecedent_generalization*	$CLASS = edible, GILL\_SIZE = broad, STALK\_SHAPE = tapering \rightarrow ODOR = none$	3
t=9	$min_{cf} = 60\%$ + *antecedent_generalization*	$GILL\_SIZE = broad, STALK\_SHAPE = tapering \rightarrow CLASS = edible$	2
t=10	*forward_chaining*	$STALK\_SHAPE = tapering \rightarrow GILL\_SIZE = broad$	8
t=11	*exception_specialization*	$GILL\_SIZE = broad, STALK\_SHAPE = tapering \rightarrow CLASS = edible$	4
t=12	*forward_chaining*	$GILL\_SIZE = broad \rightarrow CLASS = edible$	3
t=13	*forward_chaining*	$CLASS = edible, GILL\_SIZE = broad \rightarrow ODOR = none$	4
t=14	*forward_chaining*	$CLASS = edible, GILL\_SIZE = broad, ODOR = none \rightarrow STALK\_SHAPE = tapering$	10

Table 3. Exploration scenario for the MUSHROOMS dataset

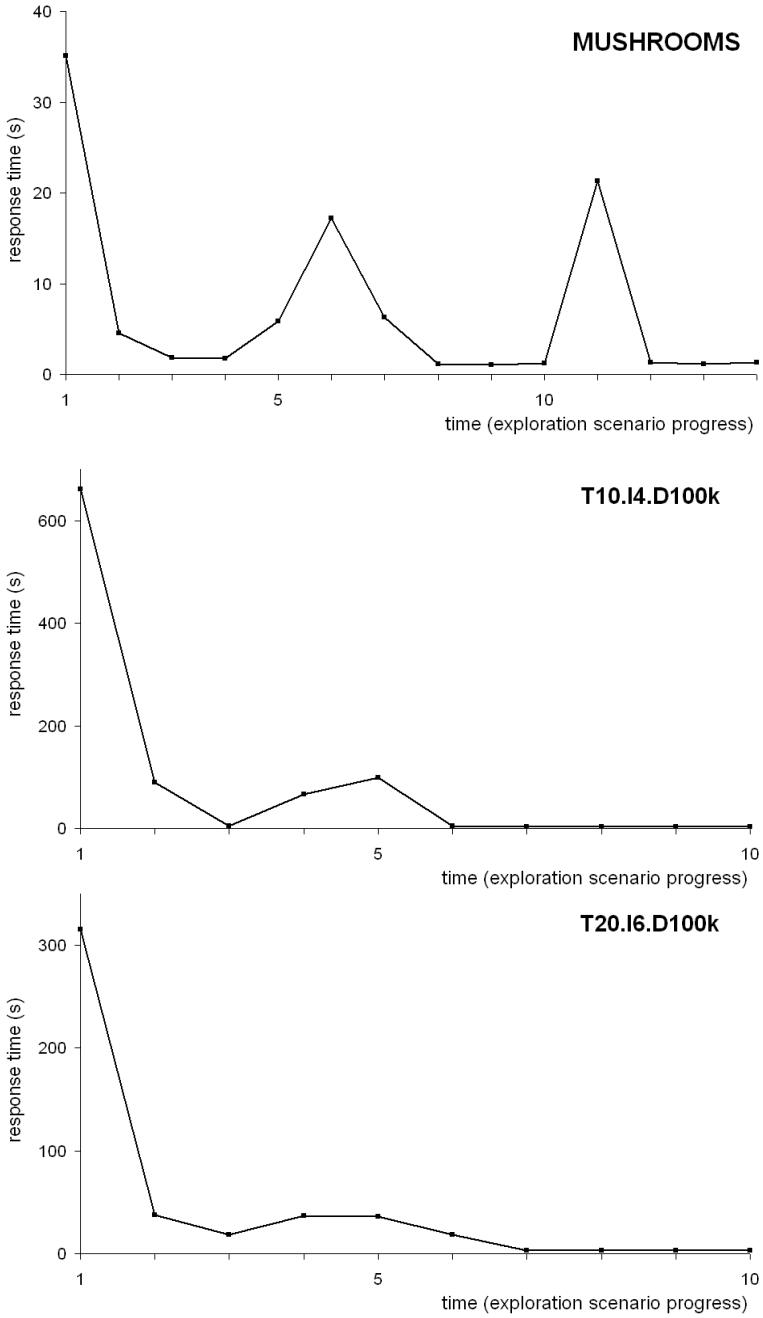


Fig. 12. Response times obtained on three exploration scenarios with *ARVis*

sta	stability	eas	motivation for easiness
fis	fighting spirit	pro	motivation for protection
ext	extroversion	rea	motivation for realization
que	questioning	mem	motivation for membership
wil	willpower	pow	motivation for power
rec	receptiveness	imp	improvisation
rig	rigor	dyn	dynamism
inc	intellectual conformism	com	communication
anx	anxiety	aff	affirmation
soc	spirit of conciliation	ind	independence

Table 4. Meaning of the attributes

7. Example of rule exploration with *ARVis*

The example presented in this section comes from a case study made with the firm PerformanSe SA on human resource management data. The data are a set of workers' psychological profiles used to calibrate decision support systems. It contains around 4000 individuals described by 20 categorical attributes (table 4). Each attribute has three possible values: "+", "0", and "-". In the example, since flashing objects cannot be seen on the figures, a transitional rule is represented in the worlds by an object with a white sphere.

The user begins by studying people that are extrovert and motivated by power. By means of the exploration initialization interface, he displays the world that contains the rules with the itemset $\{ext{=}+,\ pow{=}+\}$ in the antecedent (figure 13.A). The user explores the world. There are three rules with high confidence and high entropic implication intensity at the bottom of the arena, and one of them especially interests the user: $\{ext{=}+,\ pow{=}+\} \rightarrow \{rec{=}-\}$. To know more characteristics of this not very receptive population, he applies the neighborhood relation *forward_chaining* on this rule. The newly displayed world contains the rules with $\{ext{=}+,\ pow{=}+,\ rec{=}-\}$ in the antecedent (figure 13.B). The user finds a rule which he thinks very pertinent: $\{ext{=}+,\ pow{=}+,\ rec{=}-\} \rightarrow \{mem{=}+\}$. To know the other rules verified by these extrovert, not very receptive, and motivated by power and membership people, he applies the neighborhood relation *same_items* on the rule. In the new world, the user sees the four rules that can be built with the four items (figure 13.C). One is the transitional rule, two others are bad rules, and the fourth is a little better than the transitional rule: this is $\{ext{=}+,\ mem{=}+,\ rec{=}-\} \rightarrow \{pow{=}+\}$. To know whether all the items in the antecedent are useful in this rule, he applies the relation *generalization* on it. In the new world (figure 13.D), there is the rule $\{ext{=}+,\ mem{=}+\} \rightarrow \{pow{=}+\}$ that is as good as the transitional rule, which means that the item $\{rec{=}-\}$ was superfluous. Next, the user continues his exploration by examining the exceptions of the rule (figure 13.E).

For another exploration, the user is interested in rigorous people. He starts with the world containing the rules with $\{rig{=}+\}$ in the antecedent (figure 14.A). His attention is drawn by the rule: $\{rig{=}+\} \rightarrow \{anx{=}+\}$. This is quite a good rule, but he wants to verify whether other characteristics could better predict strong anxiety. To do so, he applies the neighborhood relation *same_consequent*. The new world contains the rules that conclude on $\{anx{=}+\}$ and shows that there is no better rule than $\{rig{=}+\} \rightarrow \{anx{=}+\}$ (figure 14.B). So the user comes back to the previous world and applies the relation *agreement_specialization* on $\{rig{=}+\} \rightarrow \{anx{=}+\}$ to know whether an additional item could improve strong

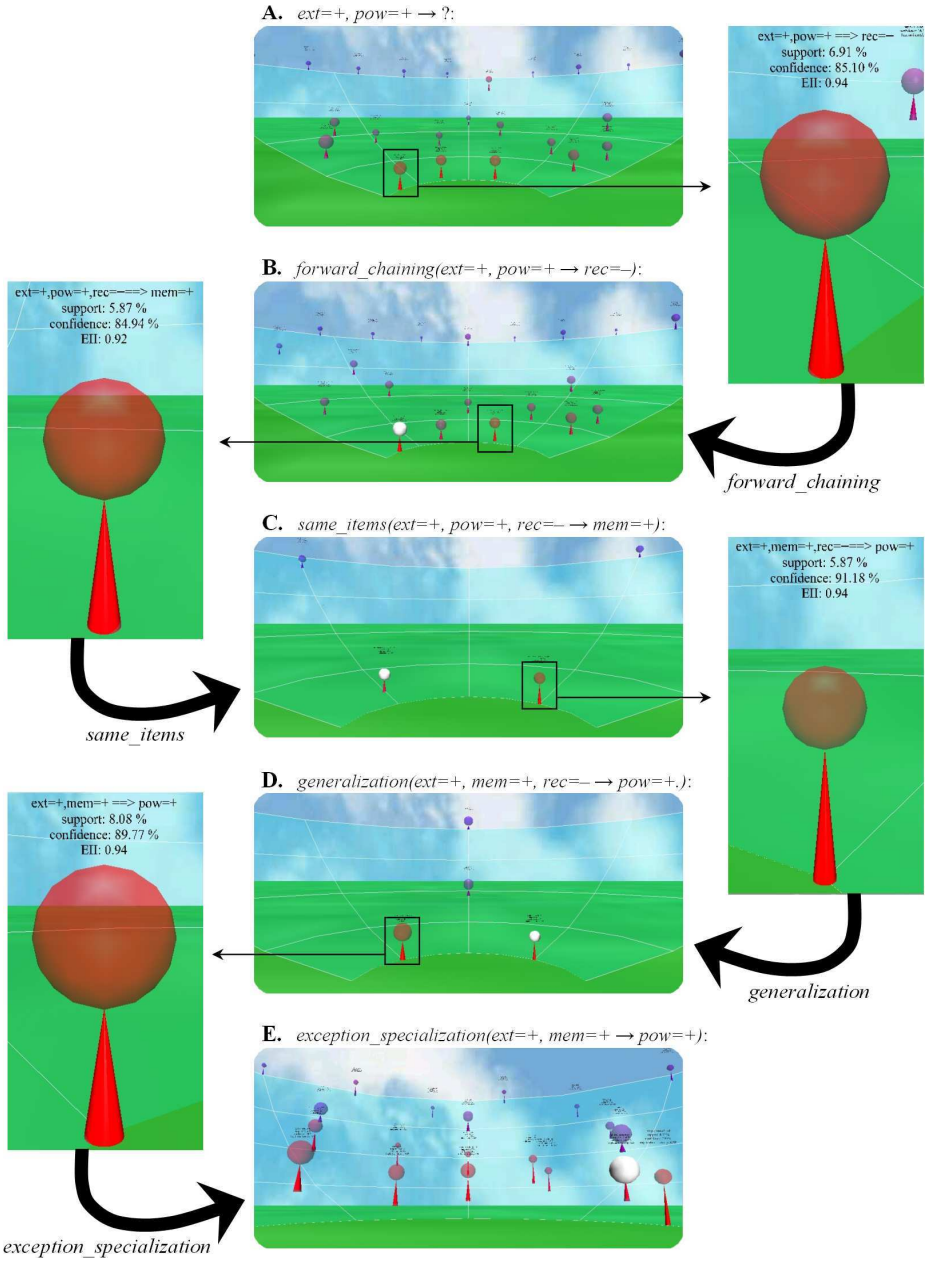


Fig. 13. An exploration with *ARVis*

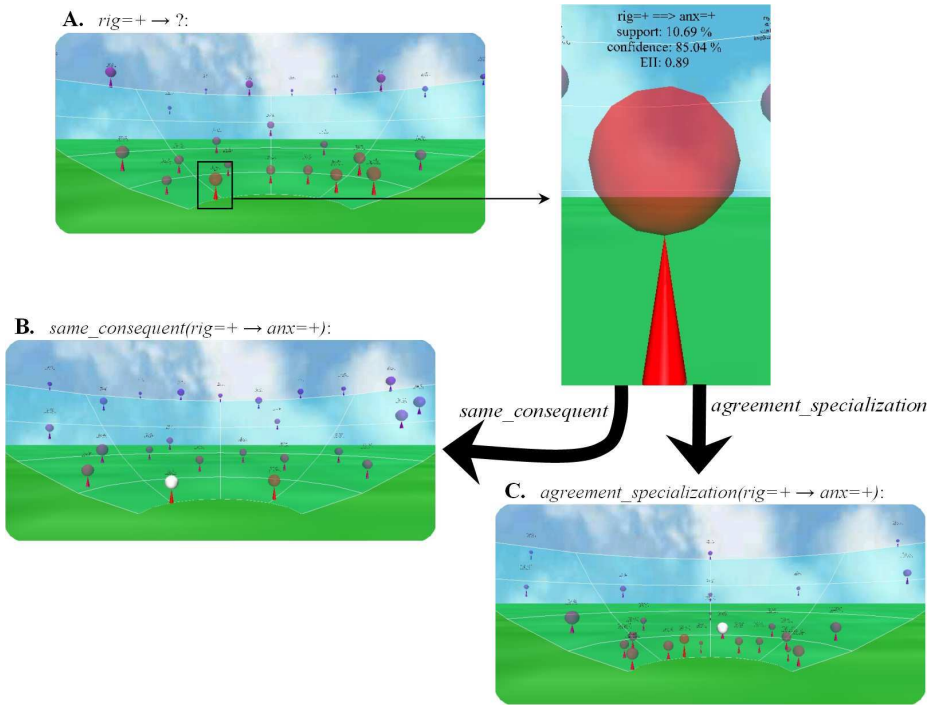


Fig. 14. An exploration with *ARVis*

anxiety prediction. The new world presents some better rules effectively (figure 14.C).

8. Conclusion

In this article, we have presented the *rule focusing* methodology for the post-processing of association rules. It enables the user to explore large sets of rules freely by focusing his/her attention on interesting subsets. The methodology relies on:

- rule interestingness measures which filter and sort the rules,
- a visual representation which speeds up comprehension and makes the comparisons among the rules easier,
- several *neighborhood relations* which connect the rules among them and underlie the interactive navigation among the rules.

We have also presented the prototype system *ARVis* which implements the rule focusing methodology by means of a 3D representation, of neighborhood relations meaningful for the user, and of a specific constraint-based rule-mining algorithm. *ARVis* takes advantage not only of the rule syntax, used in the neighborhood relations, but also of the interestingness measures, highlighted in the

representation. This dual approach is original compared to the other rule visualization methods. Moreover, *ARVis* generates the rules dynamically along the exploration by the user. Thus, the user's guidance during association rule post-processing is also exploited during association rule mining to reduce the search space and avoid generating huge amounts of rules.

The experiments we made with *ARVis* have pointed out that the tool can, on the one hand, strengthen the user in certain hypotheses and, on the other hand, provide the user with new ideas. In particular, lots of unknown rules that the user meets along the exploration arouse his/her curiosity and influence the rest of the navigation. Our future works will mainly consist in developing additional neighborhood relations among the rules. For example, a project we have with the French medical research center INSERM on cardiac pathology data requires neighborhood relations which rely on item hierarchies and generate rules with multi-item consequents. Moreover, we think that the analysis of the exploration logs of *ARVis* should reveal some "patterns of navigation" useful to create new neighborhood relations.

References

Aggarwal C.C. (2002) Towards effective and interpretable data mining by visual interaction. SIGKDD Explorations, ACM Press, vol. 3, num. 2, pp 11–22

Agrawal R., Imielinski T., and Swami A. (1993) Mining association rules between sets of items in large databases. In Proc. of the 1993 ACM SIGMOD international conference on management of data, ACM Press, pp 207–216

Agrawal R. and Srikant R. (1994) Fast algorithms for mining association rules. In Proc. of the 20th international conference on very large data bases (VLDB), Morgan Kaufmann, pp 487–499

Agrawal R., Arning A., Bollinger T., Mehta M., Shafer J., and Srikant R. (1996) The Quest data mining system. In Proc. of the 2nd ACM SIGKDD international conference on knowledge discovery and data mining, AAAI Press, pp 244-249, www.almaden.ibm.com/software/quest/

Ammoura A., Zaiane O.R., and Ji Y. (2001) Immersed Visual Data Mining: walking the walk. In BNCOD 18: Proc. of the 18th British National Conference on Databases, Springer-Verlag, pp 202–218

Andrews K. (1995) Visualising cyberspace: information visualisation in the Harmony internet browser. In Proc. of the 1995 IEEE symposium on Information Visualization, IEEE Computer Society, pp 97–104

Baird J. C. (1970) Psychophysical Analysis of Visual Space. Pergamon Press

Barthelemy J.-P. and Mullet E. (1992) A model of selection by aspects. Acta Psychologica, Elsevier Science Publishers, vol. 79, num. 1, pp 1–19

Bayardo R.J. Jr. and Agrawal R. (1999) Mining the most interesting rules. In Proc. of the 5th ACM SIGKDD international conference on knowledge discovery and data mining, ACM Press, pp 145–154

Bertin J. (1967) Smiologie Graphique. Gauthier-Villars, English translation by Berg W. J. as Semiology of Graphics (1983), University of Wisconsin Press

Bhandari I. (1994) Attribute focusing: machine-assisted knowledge discovery applied to software production process control. Knowledge Acquisition, Academic Press Ltd., vol. 6, num. 3, pp 271–294

Bisdorff R. (editor) (2003) Proc. of the mini-EURO conference on human centered processes HCP'2003. University of Luxembourg

Blake C.L. and Merz C.J. (1998) UCI Repository of machine learning databases. Irvine, CA: University of California, Department of Information and Computer Science. www.ics.uci.edu/˜mlearn/MLRepository.html

Blanchard J., Kuntz P., Guillet F., and Gras R. (2003) Implication intensity: from the basic statistical definition to the entropic version. In Bozdogan H. (editor.), Statistical data mining and knowledge discovery, Chapman & Hall/CRC Press, pp 473–485

Blanchard J. (2005) A visualization system for interactive mining, assessment, and exploration of association rules. PhD thesis, University of Nantes (in French)

Blanchard J., Guillet F., Briand H., and Gras R. (2005) Assessing rule interestingness with a probabilistic measure of deviation from equilibrium. In Proc. of the 11th international symposium on applied stochastic models and data analysis ASMDA-2005, ENST, pp 191–200

Blanchard J., Guillet F., Briand H., and Gras R. (2005) Using information-theoretic measures to assess association rule interestingness. In Proc. of the 5th IEEE international conference on data mining ICDM'05, IEEE Computer Society, pp 66–73

Bonchi F., Giannotti F., Mazzanti A., Pedreschi D. (2005) Efficient breadth-first mining of frequent pattern with monotone constraints. Knowledge and Information Systems, Springer-Verlag, vol. 8, num. 2, pp 131–153

Botta M., Boulicaut J.F., Masson C., and Meo R. (2002) A comparison between query languages for the extraction of association rules. In Proc. of the 4th international conference on data warehousing and knowlege discovery (DaWaK 2002), Springer-Verlag, Lecture Notes in Computer Science 2454, pp 1–10

Brachman, J.R., and Anand, T. (1996) The process of knowledge discovery in databases: a human-centered approach. In Fayyad U.M., Piatetsky-Shapiro G., Smyth P., Uthurusamy R. (editors), Advances in knowledge discovery and data mining, AAAI/MIT Press, pp 37–58

Braga D., Campi A., Klemettinen M., and Lanzi P.L. (2002) Mining association rules from XML Data. In Proc. of the 4th international conference on data warehousing and knowlege discovery (DaWaK 2002), Springer-Verlag, Lecture Notes in Computer Science 2454, pp 21–30

Brin S., Motwani R., Ullman J.D., and Tsur S. (1997) Dynamic itemset counting and implication rules for market basket data. SIGMOD Record, ACM Press, vol. 26, num. 2, pp 255–264

Brunk C., Kelly J., and Kohavi R. (1997) MineSet: An integrated system for data mining. In Proc. of the 3rd ACM SIGKDD international conference on knowledge discovery and data mining, AAAI Press, pp 135-138

Card S.K., Mackinlay J.D., and Schneiderman B. (editors) (1999) Readings in Information Visualization: Using vision to think. Morgan Kaufmann

Carswell C.M., Frankenberger S., and Bernhard D. (1991) Graphing in depth: perspectives on the use of three-dimensional graphs to represent lower-dimensional data. Behaviour and Information Technology, vol. 10, num. 6, pp 459–474

Chen C. (2004) Information Visualization: beyond the horizon. Springer-Verlag

Cleveland W.S. and McGill R. (1984) Graphical perception: theory, experimentation, and application to the development of graphical methods. Journal of the American Statistical Association, vol. 79, num. 387, pp 531–554

Cockburn A. and McKenzie B. (2001) 3D or not 3D? Evaluating the effect of the third dimension in a document management system. In CHI'01: Proc. of the SIGCHI conference on Human factors in computing systems, ACM Press, pp 434–441

Fayyad U.M., Piatetsky-Shapiro G., and Smyth P. (1996) From data mining to knowledge discovery: an overview. In Fayyad U.M., Piatetsky-Shapiro G., Smyth P., Uthurusamy R. (editors.), Advances in knowledge discovery and data mining, AAAI/MIT Press, pp 1–34

Fayyad U.M., Grinstein G.G., and Wierse A. (2001) Information visualization in data mining and knowledge discovery. Morgan Kaufmann

Freitas A.A. (1998) On objective measures of rule surprisingness. In Proc. of the 2nd European conference on principles of data mining and knowledge discovery (PKDD'98), Springer-Verlag, L.N.A.I. 1510, pp 1–9

Fukuda T., Morimoto Y., Morishita S., and Tokuyama T. (2001) Data mining with optimized two-dimensional association rules. ACM Transactions on Database Systems, ACM Press, vol. 26, num. 2, pp 179–213

Fule P. and Roddick J. F. (2004) Experiences in building a tool for navigating association rule result sets. In Hogan J., Montague P., Purvis M., Steketee C. (editors), CRPIT'04: Proc. of the second Australasian workshop on data mining and web intelligence, Australian Computer Society Inc., pp 103–108

Goethals B. and Van den Bussche J. (2000) On Supporting interactive association rule mining. In Proc. of the 2nd international conference on data warehousing and knowledge discovery (DaWaK), Springer-Verlag, L.N.C.S. 1874, pp 307–316

Grahne G., Lakshmanan L.V.S., and Wang X. (2000) Efficient mining of constrained correlated

sets. In Proc. of the sixteenth international conference on data engineering (ICDE), IEEE Computer Society, pp 512–521

Gras R. (1996) L'implication statistique : nouvelle mthode exploratoire de donnes. La Pense Sauvage Editions (in French)

Guillaume S., Guillet F., and Philippe J. (1998) Improving the discovery of association rules with intensity of implication. In Proc. of the 2nd European conference on principles of data mining and knowledge discovery (PKDD'98), Springer-Verlag, L.N.A.I. 1510, pp 318–327

Han J., Fu Y., Wang W., Koperski K., and Zaiane O. (1996) DMQL: a data mining query language for relational databases. In Proc. of the 1996 SIGMOD workshop on research issues on data mining and knowledge discovery (DMKD)

Han J., Chiang J.Y., Chee S., Chen J., Chen Q., Cheng S., Gong W., Kamber M., Koperski K., Liu G., Lu Y., Stefanovic N., Winstone L., Xia B., Zaiane O.R., Zhang S., and Zhu H. (1997) DBMiner: A system for data mining in relational databases and data warehouses In Proc. of CASCON'97: Meeting of Minds, pp 249–260

Han J., Pei J., and Yin Y. (2000) Mining frequent patterns without candidate generation. In Proc. of the ACM SIGMOD international conference on management of data, ACM Press, pp 1–12

Han J., An A., and Cercone N. (2000) CViz: an interactive visualization system for rule induction. In AI'00: Proc. of the 13th Biennial Conference of the Canadian Society on Computational Studies of Intelligence, Springer-Verlag, pp 214–226

Han J., Hu X., and Cercone N. (2003) A visualization model of interactive knowledge discovery systems and its implementations. Information Visualization, vol. 2, num. 2, Palgrave Macmillan, pp 105–125

Hao M.C., Dayal U., Hsu M., Sprenger T., and Gross M.H. (2001) Visualization of directed associations in e-commerce transaction data. In Proc. of VisSym 2001, pp 185–192

Hipp J., Gntzer U., and Nakhaeizadeh G. (2000) Algorithms for association rule mining - A general survey and comparison. SIGKDD Explorations, ACM Press, vol. 2, num. 1, pp 58–64

Hipp J. and Gntzer U. (2002) Is pushing constraints deeply into the mining algorithms really what we want? An alternative approach for association rule mining. SIGKDD Explorations, ACM Press, vol. 4, num. 1, pp 50–55

Hofmann H., Siebes A.P., and Wilhelm A.F. (2000) Visualizing association rules with interactive mosaic plots. In Proc. of the 6th ACM SIGKDD international conference on knowledge discovery and data mining, ACM Press, pp 227–235

Hofmann H. and Wilhelm A. (2001) Visual comparison of association rules. Computational Statistics, Physica-Verlag, vol. 16, num. 3, pp 399–415

Holland J.H. , Holyoak K.J., Nisbett R.E., and Thagard P.R. (1986) Induction : Processes of inference, learning and discovery. MIT Press

Hussain F., Liu H., Suzuki E., and Lu H. (2000) Exception rule mining with a relative interestingness measure. In Proc. of the 4th Pacific-Asia conference on knowledge discovery and data mining (PAKDD2000), Springer-Verlag, Lecture Notes in Computer Science 1805, pp 86–97

IBM (2006) DB2 Intelligent Miner Visualization. www.ibm.com/software/data/iminer/visualization/index.html

Imielinski T. and Mannila H. (1996) A database perspective on knowledge discovery. Communications of the ACM, ACM Press, vol. 39, num. 11, pp 58–64

Imielinski T. and Virmani A. (1999) MSQL: A query language for database mining. Data Mining and Knowledge Discovery, Kluwer Academic Publishers, vol. 3, num. 4, pp 373–408

Jeudy B. and Boulicaut J.-F. (2002) Optimization of association rule mining queries. Intelligent Data Analysis, IOS Press, vol. 6, num 4, pp 341–357

Keim D.A. (2002) Information visualization and visual data mining. IEEE Transactions on Visualization and Computer Graphics, IEEE Educational Activities Department, vol. 8, num. 1, pp 1–8

Klemettinen M., Mannila H., Ronkainen P., Toivonen H., and Verkamo A.I. (1994) Finding interesting rules from large sets of discovered association rules. In Proc. of the 3rd international conference on information and knowledge management (CIKM), ACM Press, pp 401–407

Kopanakis I. and Theodoulidis B. (2001) Visual data mining and modeling techniques. In Proc. of the KDD-2001 workshop on visual data mining

Kuntz P., Guillet F., Lehn R., and Briand H. (2000) A user-driven process for mining association rules. In Proc. of the 4th European conference on principles of data mining and knowledge discovery (PKDD-2000), Springer-Verlag, pp 483–489

Liu B., Hsu W., Wang K., and Chen S. (1999) Visually aided exploration of interesting association rules. In Proc. of the 3rd Pacific-Asia conference on knowledge discovery and data mining (PAKDD1999), Springer-Verlag, Lectures Notes in Artificial Intelligence 1574, pp 380–389

Liu B., Hsu W., Chen S., and Ma Y. (2000) Analyzing the subjective interestingness of association rules. IEEE Intelligent Systems, IEEE Educational Activities Department, vol. 15, num. 5, pp 47–55

Loevinger J. (1947) A systematic approach to the construction and evaluation of tests of ability. Psychological Monographs, vol. 61, num. 4

Ma Y., Liu B., and Wong C.K. (2000) Web for data mining: organizing and interpreting the discovered rules using the Web. SIGKDD Explorations, ACM Press, vol. 2, num. 1, pp 16–23

McEachren A. M. (1995) How Maps Work: representation, visualization, and design. The Guilford Press

Meo R., Psaila G., and Ceri S. (1998) An extension to SQL for mining association rules. Data Mining and Knowledge Discovery, Kluwer Academic Publishers, vol. 2, num. 2, pp 195–224

Montgomery H. (1983) Decision rules and the search for a dominance structure: towards a process model of decision making. In Humphreys P.C., Svenson O., Vari A. (editors.), Analysing and aiding decision processes, Amsterdam:North Holland, pp 343–369

Ng R.T., Lakshmanan L.V.S., Han J., and Pang A. (1998) Exploratory mining and pruning optimizations of constrained associations rules. In Proc. of the 1998 ACM SIGMOD international conference on management of data, ACM Press, pp 13–24

Ordonez C., Ezquerra N., Santana C.A. (2006) Constraining and summarizing association rules in medical data. Knowledge and Information Systems, Springer-Verlag, vol. 9, num. 3, pp 1–2

Padmanabhan B. and Tuzhilin A. (1999) Unexpectedness as a measure of interestingness in knowledge discovery. Decision Support Systems, Elsevier Science Publishers, vol. 27, num. 3, pp 303–318

Piatetsky-Shapiro G. (1991) Discovery, analysis, and presentation of strong rules. In Piatetsky-Shapiro G., Frawley W.J. (editors.), Knowledge discovery in databases, AAAI/MIT Press, pp 229–248

Rainsford C.P. and Roddick J.F. (2000) Visualisation of temporal interval association rules. In Proc. of the 2nd international conference on intelligent data engineering and automated learning (IDEAL 2000), Springer-Verlag, pp 91–96

Robertson G., Czerwinski M., Larson K., Robbins D.C., Thiel D., and van Dantzich M. (1998) Data mountain: using spatial memory for document management. In UIST'98: Proc. of the 11th annual ACM symposium on user interface software and technology, ACM Press, pp 153–162

SAS (2006) Enterprise Miner. www.sas.com/technologies/analytics/datamining/miner/

Sebag M. and Schoenauer M. (1988) Generation of rules with certainty and confidence factors from incomplete and incoherent learning bases. In Proc. of the European knowledge acquisition workshop EKAW'88, Gesellschaft fr Mathematik und Datenverarbeitung mbH, pp 28.1–28.20

Schneiderman B. (2002) Inventing discovery tools: combining information visualization with data mining. Information Visualization, Palgrave Macmillan, vol. 1, num. 1, pp 5–12

Silberschatz A. and Tuzhilin A. (1996) User-assisted knowledge discovery: how much should the user be involved. In Proc. of the 1996 SIGMOD workshop on research issues on data mining and knowledge discovery (DMKD)

Silberschatz A. and Tuzhilin A. (1996) What makes patterns interesting in knowledge discovery systems. IEEE Transactions on Knowledge and Data Engineering, IEEE Educational Activities Department, vol. 8, num. 6, pp 970–974

Silverstein C., Brin S., and Motwani R. (1998) Beyond market baskets: Generalizing association rules to dependence rules. Data mining and knowledge discovery, Kluwer Academic Publishers, vol. 2, num. 1, pp 39–68

Simon H.A. (1979) Models of Thought. Yale University Press

Spence I. (1990) Visual psychophysics of simple graphical elements. Journal of Experimental Psychology: Human Perception and Performance, vol. 16, num. 4, pp 683–692

Spence R. (2000) Information Visualization. Addison Wesley

Srikant R., Vu Q., and Agrawal R. (1997) Mining association rules with item constraints. In Proc. of the 3rd ACM SIGKDD international conference on knowledge discovery and data mining, AAAI Press, pp 67–73

Suzuki E. (2002) Undirected discovery of interesting exception rules. International Journal of Pattern Recognition and Artificial Intelligence, World Scientific Publishing Company, vol. 16, num. 8, pp 1065–1086

Tan P.-N., Kumar V., and Srivastava J. (2004) Selecting the right objective measure for association analysis. Information Systems, Elsevier Science Publishers, vol. 29, num. 4, pp 293–313

Tufte E. (1983) The Visual Display of Quantitative Information. Graphics Press

Tuzhilin A. and Adomavicius G. (2002) Handling very large numbers of association rules in the analysis of microarray data. In KDD'02: Proc. of the eighth ACM SIGKDD international conference on knowledge discovery and data mining, ACM Press, pp 396–404

Unwin A. R., Hofmann H., and Bernt K. (2001) The TwoKey plot for multiple association rules control. In Proc. of 5th European conference on principle and practice of knowledge discovery in databases (PKDD'01), Springer-Verlag, pp 472–483

Ware C. and Franck G. (1996) Evaluating stereo and motion cues for visualizing information nets in three dimensions. ACM Transactions on Graphics, vol. 15, num. 2, pp 121-140

Wilkinson L. (1999) The Grammar of Graphics. Springer-Verlag

Wong P.C., Whitney P., and Thomas J. (1999) Visualizing association rules for text mining. In Proc. of the 1999 IEEE symposium on information visualization, IEEE Computer Society, pp 120–123

Author Biographies

Dr. **Julien Blanchard** earned his Ph.D. in 2005 from Nantes University (France) and is currently an assistant professor at Polytechnic School of Nantes University. He is the author of a book chapter and 7 journal and international conference papers in the areas of visualization and interestingness measures for data mining.

Assistant Prof. **Fabrice Guillet** is currently a member of the LINA laboratory (CNRS 2729) at Polytechnic graduate School of Nantes University (France). He hold a Ph.D. in computer science in 1995 from the Ecole Nationale Supérieure des Télécommunications de Bretagne. He is author of 35 international publications in data mining and knowledge management. He is a foundator and a permanent member of the Steering Committee of the annual EGC French-speaking conference.

Prof. **Henri Briand** earned his Ph.D. in 1983 from Paul Sabatier University located in Toulouse (France) and has over 100 publications in database systems and database mining. He was the head of the Computer Engineering Department at Polytechnic School of Nantes University. He was in charge of a research team in the data mining domain. He is responsible for the organization of the Data Mining Master in Nantes University.

Correspondence and offprint requests to: Julien Blanchard, École Polytechnique de l'Université de Nantes, Département Informatique, La Chantrerie, BP 50609, Nantes cedex 3, 44306, France. Email: julien.blanchard@polytech.univ-nantes.fr

Annexe C

ATHANOR : une approche pour la gestion des connaissances de maintenance sur les systèmes complexes[103]

Dans *Première journée Sstèmes d'information pour l'aide à la décision en ingénierie système, JESIADIS 2002.*

ATHANOR
Une approche de gestion de connaissances procédurales pour la maintenance de systèmes complexes

Fabrice Guillet[*], Vincent Philippé[**], Jacques Philippé[*,**], Dominique Follut[**]

*LINA - Ecole polytechnique de l'université de Nantes, rue C. Pauc, 44087 Nantes.
Fabrice.Guillet@polytech.univ-nantes.fr
**PERFORMANSE SA -Atlanpôle - La Fleuriaye BP703 44481 Carquefou.
{Vincent.Philippe, Jacques.Philippe, Dominique.Follut}@performanse.fr

Résumé. Un grand nombre d'entreprises sont confrontés à des problèmes stratégiques de gestion des connaissances, d'autant plus critiques que les connaissances portent sur des systèmes complexes, qui nécessitent la mise en œuvre d'une démarche instrumentalisée intégrant le déploiement d'une plateforme opérationnelle dans le système d'information de l'entreprise. C'est dans ce cadre que s'inscrit la démarche Athanor pour la maintenance de systèmes complexes. En nous inspirant des méthodes de capitalisation et de formalisation des connaissances pour la conception de mémoires organisationnelles, nous avons conçu un serveur de connaissances orienté processus qui implémente les services de capitalisation-évolution des connaissances, d'aide à la décision pour le diagnostic, et de formation, en conjonction avec des modèles en réalité virtuelle des machines de tri et une documentation électronique. Cet outil, fondé sur trois modèles structurés et en interrelation, utilise un support universel, Internet, pour son déploiement. Une implantation d'Athanor sur un système complexe a été réalisé à La Poste : SAMANTA (Système d'Aide à la MAiNtenance des Trieuses Automatiques).
Mots clés. gestion des connaissances, mémoire organisationnelle, aide à la décision, maintenance de systèmes complexes, serveur de connaissances, réalité virtuelle, qualité des connaissances.

1. Introduction

Un nombre croissant d'entreprises sont confrontées à des problèmes stratégiques de gestion des connaissances : érosion de leurs experts pour diverses raisons (pyramide des âges, mobilité, …), accélération des cycles technologiques, accroissement de la durée de vie des produits, et réutilisation des connaissances. Ces problèmes s'avèrent d'autant plus difficiles à résoudre que d'une part les systèmes à gérer possèdent des structures complexes (assemblages d'un très grand nombre de composants, multiplicité des technologies, des fonctions, et des utilisateurs/acteurs), que d'autre part les sources de connaissances relatives à ces systèmes sont fragmentaires et multiples (experts, techniciens, documentation techniques, fiches de maintenance, GED, bases de données…), et qu'enfin les connaissances les plus proches des cœurs de métiers sont souvent de nature procédurale (expertise

processus) et tacite. C'est dans ce contexte de passage à l'échelle sur des systèmes complexes que s'inscrit notre serveur Athanor pour la gestion des connaissances.

Les besoins exprimés sont lourds et nécessitent la mise en œuvre d'une démarche complète de gestion des connaissances, de la phase de recueil jusqu'à la phase de déploiement. Ces besoins dont les enjeux principaux ([DIEN 99]) sont la *capitalisation*, la *diffusion* et l'*innovation*, peuvent être récapitulés ainsi :

- *recueillir* les connaissances, la *formaliser*, puis la *mémoriser*, en tenant compte de la multiplicité des sources de connaissances
- *diffuser/déployer* les connaissances recueillies, en tenant compte des moyens de diffusion ;
- faciliter *l'évolution/maintenance* des connaissances afin d'assurer leur pérennité ;
- accélérer l'*appropriation* des connaissances en tenant compte des usages.
- simplifier le *transfert* de connaissances, en intégrant des fonctionnalités pédagogiques (*formation* de nouveaux experts…).

En réponse à ces besoins, une première possibilité consiste à mettre en œuvre une démarche classique de construction d'une mémoire organisationnelle de type documentaire, en s'appuyant sur une méthode de formalisation des connaissances (MKSM [ERM 96], REX , KADS [SCH 94]). Après avoir délimité les connaissances utiles, généralement celles relative au cœur de métier d'une entreprise, puis procédé au recueil des connaissances par des interviews, cette démarche aboutit à la rédaction d'un document, le livre des connaissances, capitalisant l'ensemble des connaissances relatives à un système. Une des difficultés rencontrées, réside dans le fait que le livre de connaissances a été conçu pour répondre au besoin de capitalisation des connaissances et qu'il s'avère mal adapté aux autres besoins liés à son opérationnalisation (déploiement, maintenance, transfert, …)

Une seconde possibilité, plus opérationnelle, est de concevoir une documentation technique intelligente [PHA 99] offrant un support plus évolué pour les sources documentaires techniques. L'originalité de cette approche est de franchir la passivité traditionnelle des documents, non seulement en lui choisissant un support de présentation électronique multimédia plus interactif, diffusable et maintenable, mais aussi en y couplant un système à base de connaissances capable de fournir en ligne une activité d'aide à la décision grâce à des connaissances activables (procédures de diagnostic, de maintenance, …). Bien qu'elle permette un meilleur déploiement des documents techniques, cette approche n'accorde toutefois qu'un statut secondaire à la connaissance activable et privilégie les connaissances documentaires statiques à l'instar du livre des connaissances.

Aussi, dans la prolongation des démarches précédentes, nous proposons une nouvelle approche, ATHANOR, orientée vers le déploiement et l'opérationnalisation de connaissances portant sur des systèmes complexes. L'approche ATHANOR est fondée sur un serveur de connaissances qui permet la capitalisation, la diffusion et le maintien de la connaissance sous forme multimédia et activable ([PEN 00]), et intègre des services d'aide à la décision, et de formation, en conjonction avec des modèles en réalité virtuelle des systèmes et une documentation électronique nécessaire dans le cadre des systèmes complexes.

Dans cet article, nous présenterons les caractéristiques du serveur de connaissances ATHANOR à travers les trois modèles couplés qu'il incorpore, ainsi que son architecture modulaire dont chaque module offre une vue dédiée sur les modèles. Nous détaillerons plus particulièrement le modèle processus et la représentation graphique qui lui est associée. Puis nous présenterons deux modules entretenant de fortes relations avec les trois modèles et en

association avec les représentations en réalité virtuelles et la documentation électronique : le module Expert qui permet l'édition et le maintien des connaissances, et le module Praticien activant les connaissances en mode résolution de problème pour l'aide à la décision. Enfin nous montrerons l'intérêt du module Manager permettant, à travers des indices numériques, de surveiller la qualité des connaissances décrites dans les modèles.

2. Les modèles de connaissances d'ATHANOR

Le serveur de connaissances ATHANOR est fondé sur trois modèles principaux permettant de décrire les connaissances selon trois points de vue complémentaires : un modèle processus orienté vers les processus métiers (comment), un modèle organique du système supportant les processus (quoi), et un double modèle compétences/organigramme (qui). A un niveau plus global, ces trois modèles entretiennent de fortes interrelations afin d'exprimer la richesse multidimensionnelle des connaissances : les connaissances concernent des processus portant sur des composants d'un système dont la manipulation nécessite des compétences mises en oeuvre par des individus (Fig 1).

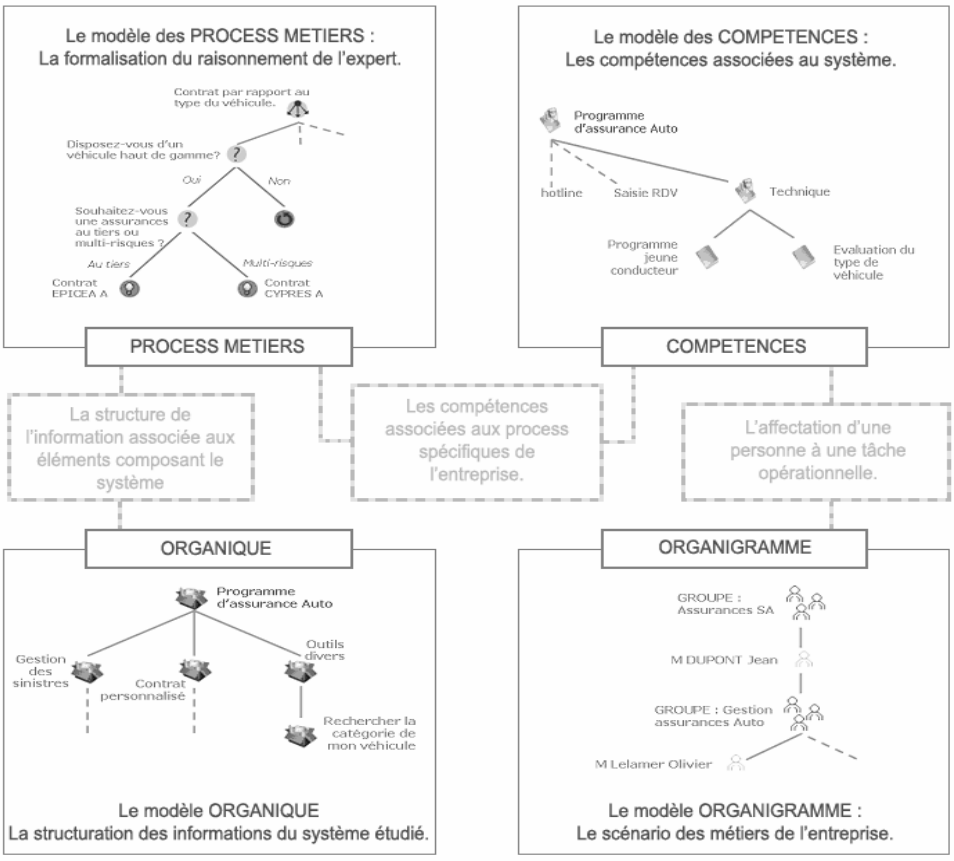


Fig. 1 – Les modèles et leurs interactions

Plus précisément, le modèle processus a pour objet la description de processus métier pratiqués par des experts. Il permet de maintenir des connaissances de nature procédurale qui sont décomposées en un enchaînement d'étapes. Chaque étape y est décrite en fonction de ses propriétés, et l'enchaînement des étapes traduit des règles de raisonnement. Le modèle organique permet de décrire la structure du système à travers sa décomposition en une hiérarchie de composants et de groupe de composants. Enfin, le double modèle compétences/organigramme permet de décrire d'une part une hiérarchie de compétences requises pour mener les étapes du processus, et d'autre part la hiérarchie des acteurs et de leurs compétences acquises sur le système. Chaque compétence est définie en terme de savoir, savoir-faire et savoir-être.

Dans un souci de simplification de l'accès à cet ensemble de modèles, ceux-ci sont conçus afin de disposer d'une représentation graphique canonique : un arbre ou un graphe.

3. Le Serveur de Connaissances ATHANOR

D'un point de vue plus technique, le serveur de connaissances ATHANOR est conçu de manière modulaire et extensible, et s'appuie sur un support technique universel, Internet, facilitant le déploiement des connaissances. Plus précisément, l'architecture technique comporte en son cœur un serveur de connaissances gérant une base de connaissances, et en périphérie un ensemble de modules graphiques proposant différentes vues fonctionnelles sur les modèles (cf. fig. 2) :

- Un module *Praticien* pour activer les connaissances en mode résolution de problème,
- Un module *Expert* pour décrire et mettre à jour les modèles et leurs associations,
- Un module *Manager* pour surveiller avec un tableau de bord la base de connaissance et son utilisation,
- Un module *Pédagogue* pour consulter les connaissances sous forme graphique commentée.

En complément, un ensemble de modules additionnels sont proposés pour faciliter à la fois l'intégration d'ATHANOR dans un système d'information préexistant et y promouvoir son utilisation :

- Un module *réalité virtuelle* en liaison avec le modèle organique, pour l'utilisation de représentations en trois dimensions du système supportant les connaissances,
- Un module de *communication*, qui permet aux utilisateurs s'échanger des informations,
- Un module de *gestion de groupes*, qui permet gérer les utilisateurs et leurs droits d'accès,
- Un module d'*expertise nomade*, générant une version allégée des modèles pour des périphériques tel qu'un PDA,
- Un module d'*accès au système de gestion documentaire du SI*, permettant le référencement de documents externes dans la base de connaissances.

La gestion de la complexité des systèmes sur lesquels porte les connaissances est facilitée par l'utilisation des représentations graphiques associées aux trois modèles, afin d'assurer la lisibilité et l'appropriation des connaissances. Les modules Expert et Pédagogue permettent d'organiser les connaissances procédurales, d'appréhender la complexité structurelle du système, et de gérer les compétences associées à ces modèles. Le module Manager, sorte de

tableau de bord sur la base de connaissances, a deux objectifs : l'un est d'offrir une synthèse sur la structure de la base elle-même, c'est l'étude de la structure *statique* des connaissances, et l'autre, considère son aspect *dynamique*, la manière dont elle évolue au fur et à mesure des utilisations. Le module Praticien est accessible à tous les utilisateurs. Il permet d'actionner la base de connaissances, en déroulant les processus métiers par un jeu de questionnement, et suivant une problématique de résolution de problème.

Chacun de ces modules possède sa propre interface graphique avec l'utilisateur et est conçu comme un client séparé fonctionnant dans un navigateur web, s'appuyant sur les technologies de l'Internet (dhtml, java, …) et communiquant par requêtes avec le noyau Apache/Prolog du serveur de connaissances. En interne, le serveur de connaissances maintient les modèles de connaissances dans des bases de connaissances opérationnelles implémentées en prolog. Cette implémentation en prolog est transparente pour les utilisateurs et est totalement cachée par les interfaces graphiques des différents modules. Ce choix du langage Prolog pour l'implémentation du serveur a l'avantage de faciliter la gestion interne des connaissances recueillies, mais surtout de réaliser un stockage opérationnel des connaissances permettant leur activation à travers des moteurs d'inférence en mode résolution de problèmes. Il offre également de grandes perspectives d'extension du modèle actuel de connaissances.

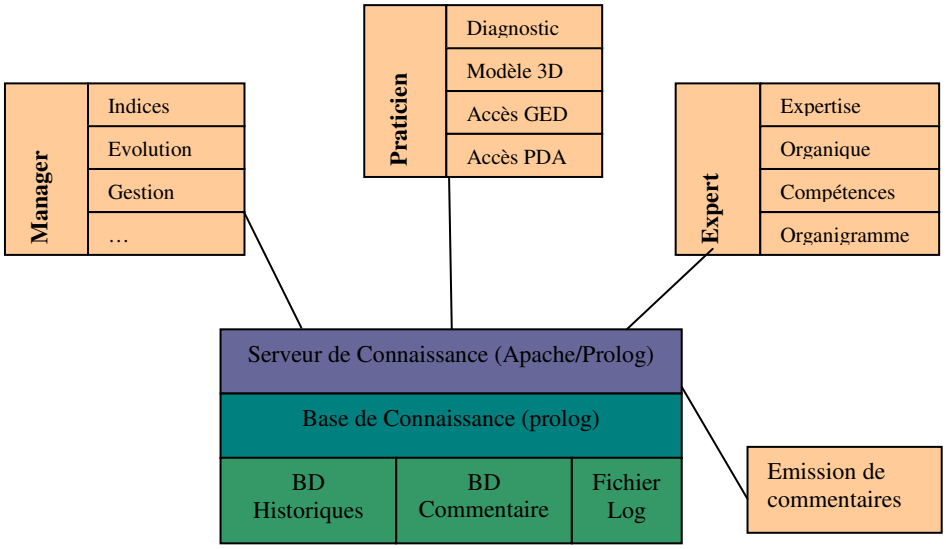


Fig. 2 – Architecture Modulaire du Serveur de Connaissances

Enfin, en réponse au besoin de *diffusion des connaissances*, le choix Intranet/Apache facilite l'accès au serveur de connaissances sur un réseau d'entreprise. Il permettra aussi d'en observer l'utilisation par analyse des fichiers log, à l'aide de techniques de découverte de connaissances.

4. La Formalisation Graphique des Connaissances : le modèle processus

Ainsi que les modèles *organique* et *compétences/organigramme*, le modèle *processus* dispose d'une représentation graphique canonique : un graphe. Ce modèle permet de représenter des connaissances procédurales liées à un savoir-faire sous la forme d'un processus décomposé en une suite d'étapes à mener. Ainsi, dans le cas d'un diagnostic, chaque étape consiste à tester des hypothèses sur l'état des composants ou des fonctionnalités du système. L'enchaînement des étapes suit une logique d'efficacité : des hypothèses les plus simples aux plus complexes.

Le graphe du processus est appelé *logigramme* (Fig 3), et le raisonnement de l'expert s'y traduit par un parcours du graphe depuis sa racine (de haut en bas et de gauche à droite). Chaque étape du raisonnement correspondant à un sommet de ce graphe. Cette représentation graphique peut être vue comme une généralisation des arbres de décision et des arbres de défaillance, les enrichissant de sommets structurants afin de prendre en compte le modèle de raisonnement des experts.

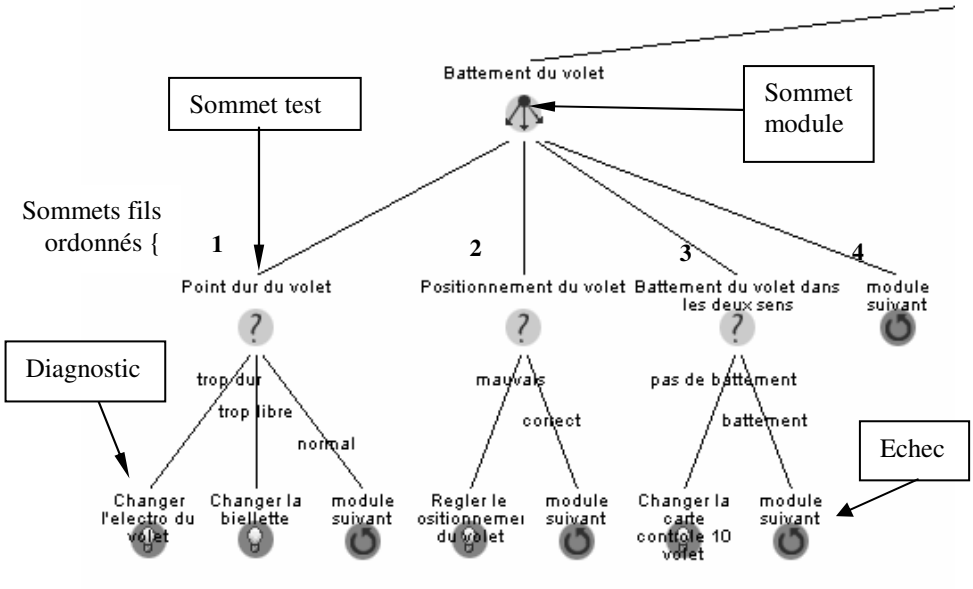


Fig. 3 – Représentation par logigramme du modèle processus

En complément de cette logique d'enchaînement portée par la structure globale du logigramme, nous avons défini quatre types de sommets dont la sémantique diffère. Les deux premiers types sont des sommets intermédiaires structurants :

- Les sommets *test* associés à une variable, typiques des arbres de décision, dont les fils ne sont pas ordonnés, mais dont chaque arc est associé à une valeur de la variable. La variable est généralement associée à l'état de fonctionnement dans lequel se trouve un élément du système à diagnostiquer.

- Les sommets *module*, absents des arbres de décision et de défaillance, dont les fils sont ordonnés de gauche à droite et généralement du plus simple au plus complexe, au sens de l'expert. Chacun de ces sommets permet de définir un module de connaissances.

Les sommets *module* permettent ainsi d'intégrer des principes cognitifs caractéristiques des stratégies de décision expertes [BAR 92], dont un principe de parcimonie/décidabilité :

- Les premiers sommets fils d'un module permettent d'arriver à une décision à moindre coût par des opérations simples (parcimonie)
- Les sommets fils suivant offrent la possibilité de réaliser des opérations de plus en plus complexe afin d'arriver à une prise de décision même si elle s'avère coûteuse (décidabilité).

Nous proposons aussi deux autres types pour les sommets terminaux (feuilles ou puits) :

- Les sommets *diagnostic*, indiquant la fin du processus : résolution du problème et la réparation à effectuer.
- Les sommets associés à un *échec* provoquant la mise en œuvre d'un mécanisme de retour par back-tracking au dernier sommet module traité et la transition au sommet suivant au sens de l'ordre induit par ce sommet module.

Enfin, chaque sommet cette représentation est décrit par des informations propres :

(a) le composant suspecté : élément supposé défectueux à ce stade du raisonnement (composant décrit dans le modèle organique)

(b) L'accès a une représentation en réalité virtuelle du composant suspecté afin de faciliter sa localisation (associé au modèle organique).

(c) La description multimédia du mode opératoire à suivre pouvant inclure des liaisons vers une documentation complémentaire, des images, des vidéos, des sons, des modèles en réalité virtuelle, des odeurs, ainsi que des informations connexes du type instument à utiliser pour évaluer l'état du ou des composants en cause.

(d) La liste des compétences requises pour opérer cette intervention (associé au modèle compétences) et les personnes de l'organisation qui possèdent ces compétences.

Une propriété importante de cette formalisation graphique, réside dans la possibilité de la transformer en un ensemble de règles de production, en traduisant l'ensemble des chemins menant de la racine à chacune des feuilles. Ainsi, le logigramme (Fig 3) se transformerait en 4 règles :

Règle 1 : *si* point dur du volet = trop dur *alors* changer l'électro du volet
Règle 2 : *si* point dur du volet = trop libre *alors* changer la biellette
Règle 3 : *si* point dur du volet = normal et positionnement du volet = mauvais *alors* régler le positionnement du volet
Règle 4 : *si* point dur du volet = normal et positionnement du volet = correct et battement du volet dans les 2 sens = pas de battement *alors* changer la carte contrôle 10 volet

Cette représentation graphique des connaissances procédurales a l'avantage d'être beaucoup plus intelligible et synthétique qu'un ensemble équivalent de règles de production.

5. L'Editeur Graphique de Connaissances : le module Expert

Dans un contexte d'appropriation de l'outil par les experts et afin de faciliter l'évolution des connaissances, un éditeur spécifique a été développé : le module Expert (Fig 4).

Athanor – gestion de connaissances procédurales sur des systèmes complexes

L'éditeur comporte quatre onglets permettent de modifier graphiquement l'intégralité des informations portées par chacun des trois modèles internes et ainsi de modifier l'intégralité de la base de connaissances. Chaque modèle dispose de sa propre interface graphique et de liaisons avec les autres modèles. L'ordre des onglets propose un ordre indicatif de saisie des connaissances :

- Organiser les différents processus selon une hiérarchie de classement propre au métier (ex : symptômes initiaux pour le diagnostic) (Fig 4).
- Construire le logigramme de chaque processus, et y décrire chaque étape (Fig 5 et 6).
- Construire le modèle organique du système et décrire chacun des composants avec les informations multimédias disponibles : sons, images, vidéos, lien vers des documents externes, … (Fig 7).
- Décrire le référentiel des compétences associées en terme de savoir, savoir-faire et savoir-être (Fig 8).
- Identifier l'organigramme des individus opérant sur ce système (Fig 9).

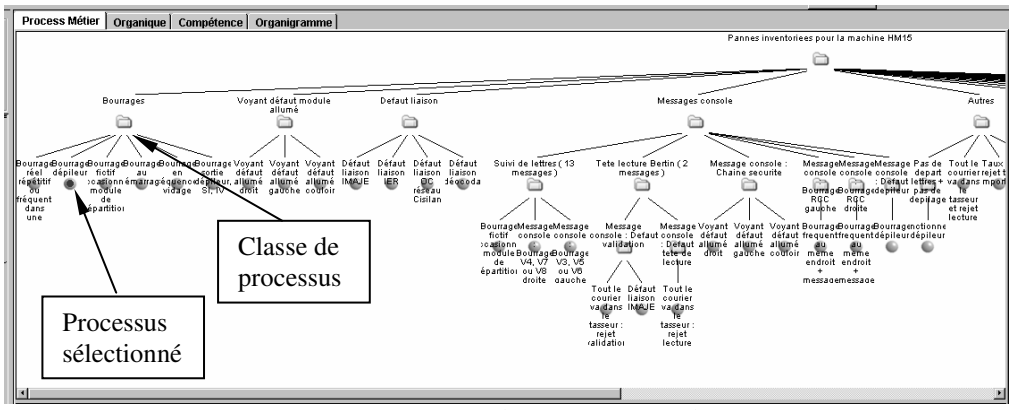


Fig. 4 – Organisation des processus en classes

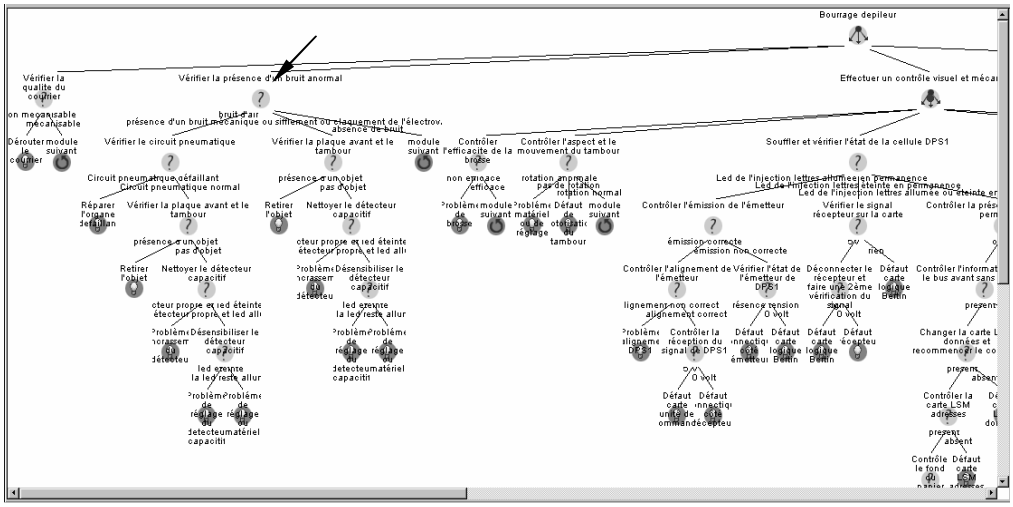


Fig. 5 – Logigramme du processus sélectionné en Fig 4

Les connaissances décrites par ce formalisme graphique sont ensuite stockées sous forme relationnelle en prédicats *prolog* dans une base de connaissances. Les connaissances pourront ensuite être déclenchées en mode résolution de problèmes pour l'aide au diagnostic à travers le module Praticien.

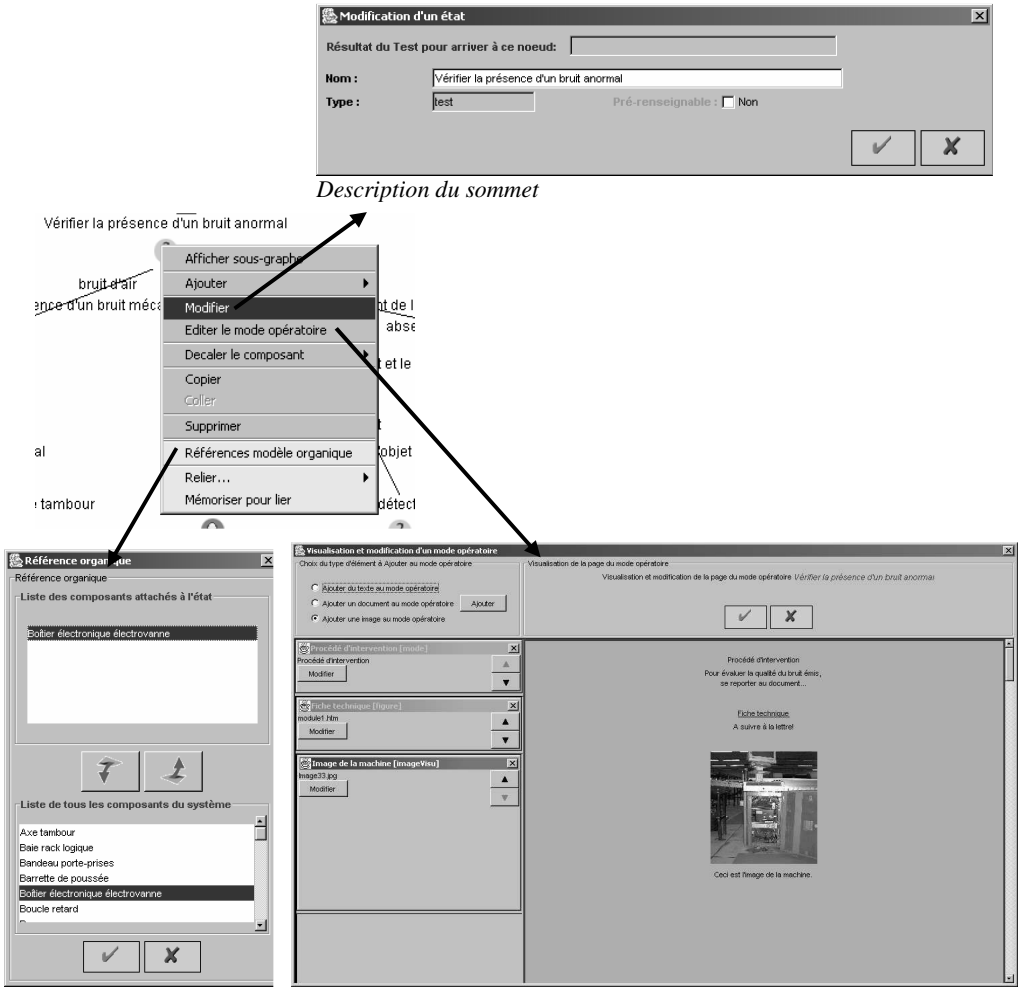
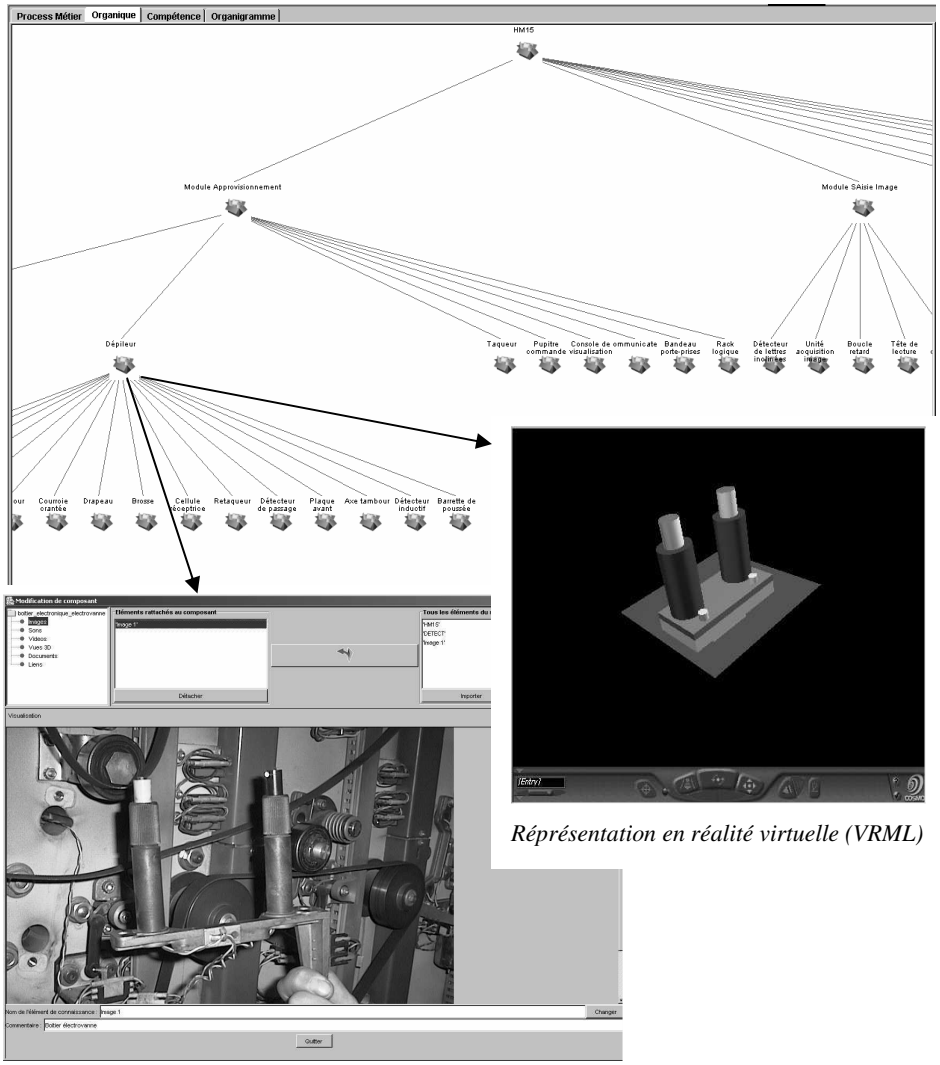


Fig. 6 – Edition des informations associées au sommet sélectionné en Fig 5



Réprésentation en réalité virtuelle (VRML)

Description du sommet

Fig. 7 – Edition du modèle organique et des informations associées à chaque sommet(composant)

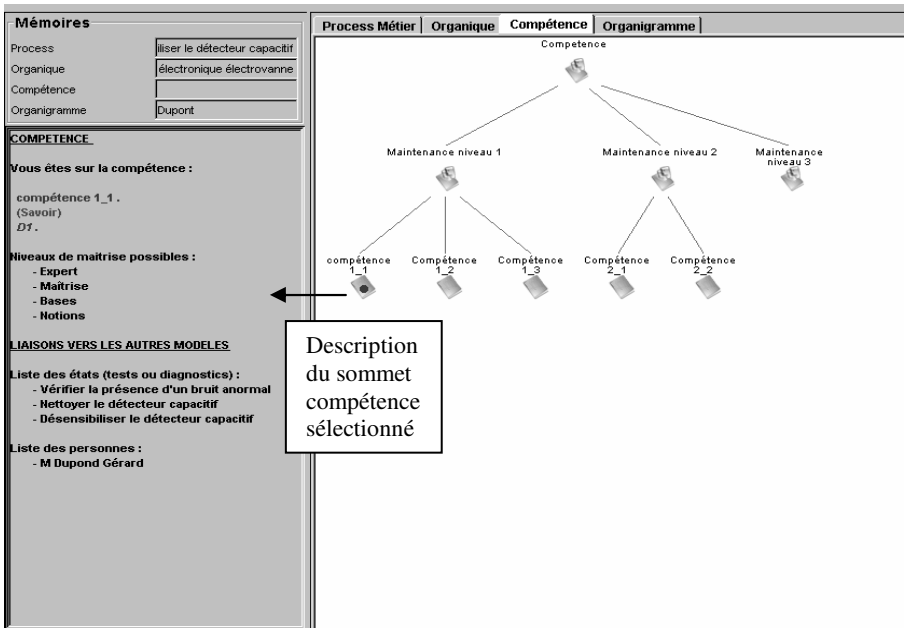


FIG. 8 – *Edition du modèle compétences*

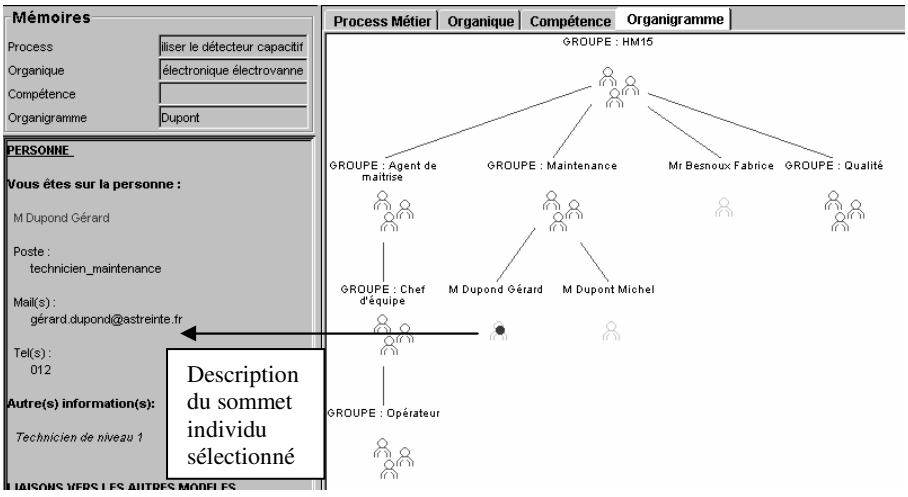


Fig. 9 – Edition du modèle organigramme

Les associations entre les modèles de connaissances sont réalisées de manière aussi interactive que la constitution des modèles eux-mêmes. Elles sont mentionnées dans la partie gauche des interfaces d'édition, pour en faciliter la lisibilité.

6. L'aide à la décision : le module Praticien

L'interface (Fig 10 et 11) permet de dérouler automatiquement en mode résolution de problème un questionnaire d'aide à la décision auprès d'un utilisateur non expert. Dans le cas d'un diagnostic, les réponses données par l'utilisateur permettent de renseigner le serveur de connaissances sur l'état des composants de la machine, afin d'affiner le diagnostic jusqu'à ce que l'origine de la panne (ou fin du processus courant) soit détectée. A chaque étape du questionnaire, l'utilisateur est guidé par une description de la question : le nom de l'opération à mener, le composant suspecté, le mode opératoire à suivre pour fournir la réponse, un historique des questions précédentes. Le questionnaire est dynamiquement généré à partir des connaissances stockées dans la base de connaissances, et est dirigé par le modèle processus : chaque question est associée à une étape du processus, et le questionnaire correspond à un cheminement dans le graphe du processus.

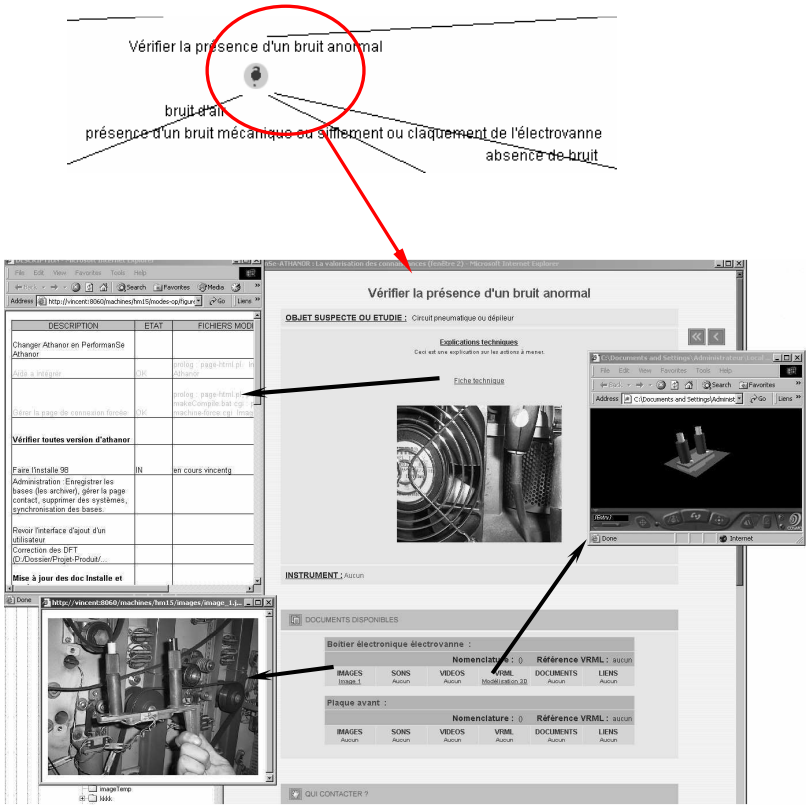


Fig. 10 – Module praticien : questionnaire et informations liées

A tout moment du questionnaire l'utilisateur a la possibilité d'obtenir des informations sur l'étape en cours décrite dans le modèle processus, mais aussi sur la partie des autres modèles en relation avec cette étape. En particulier, il peut localiser sur des représentations en réalité virtuelle la position des composants suspectés maintenus par le modèle organique

(Fig 12) ; et le modèle compétences/organigramme permet de repérer les individus disposant des compétences requises cette étape du processus (Fig 11).

Enfin, le module d'expertise garde une trace du cheminement de l'utilisateur dans la base de connaissances, en vu de son exploitation ultérieure par le profil « Manager ».

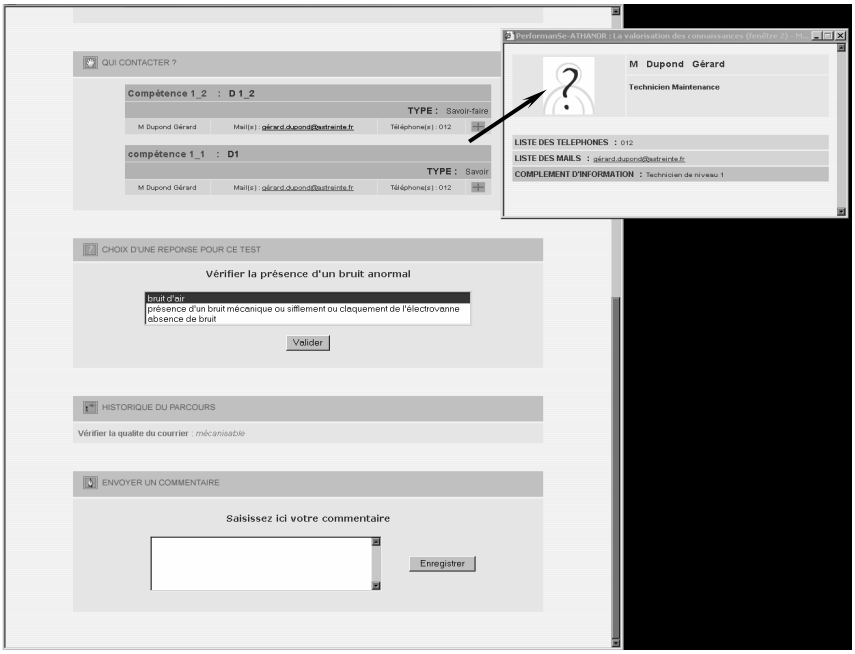


Fig. 11 – Module praticien (suite) : questionnaire et informations liées

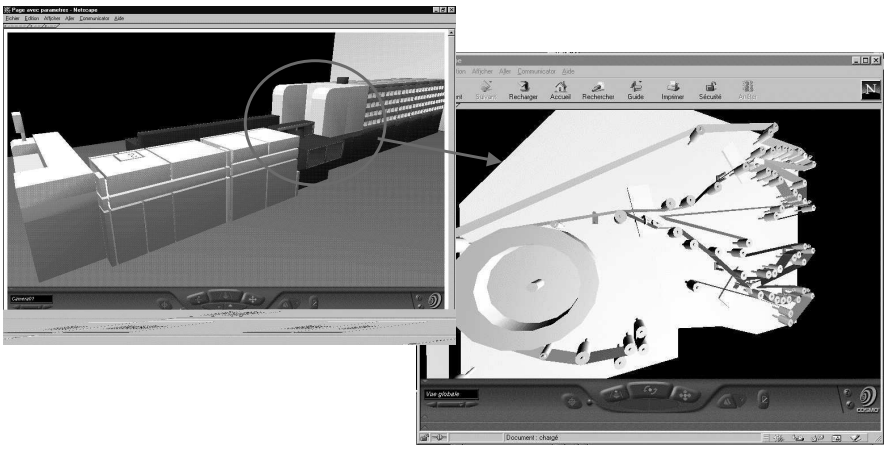


Fig. 12 – Module de réalité virtuelle – localisation de composants dans un système complexe

7. Le suivi des connaissances : le module Manager

Le module Manager récapitule dans un tableau de bord un ensemble d'indicateurs numériques permettant de surveiller la structure, l'évolution et les usages des connaissances. En effet, dans le contexte d'un système complexe, il est important de pouvoir contrôler rapidement l'adéquation entre les modèles implémentés dans Athanor et le système réel dans son contexte opérationnel.

Trois types d'information sont présentés dans ce module à travers des indices statistiques élémentaires :

- Des indices structurels statiques permettant de mesurer la richesse de la modélisation à un instant donné. Ces indices comptabilisent le nombre de sommets, le nombre d'informations associées à chaque sommet et le nombre de relations entre sommets dans les différents modèles, ainsi qu'entre les modèles.
- Des indices structurels dynamiques qui permettent de mesurer l'évolution dans le temps des indicateurs précédents.
- Des indices d'usage, mesurant l'activité des utilisateurs à travers leurs accès au serveur de connaissances.

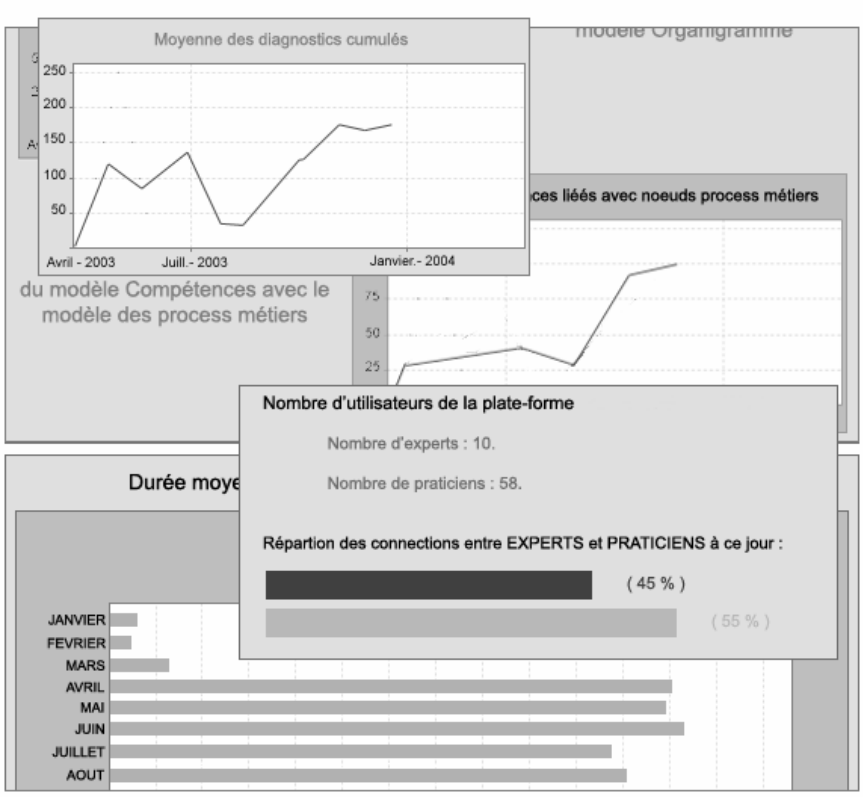


Fig. 13 – Module Manager : Présentation synthétique de quelques indicateurs

8. Conclusion

L'ensemble de la démarche ATHANOR et l'approche serveur de connaissances permet de gérer un cycle complet des connaissances, depuis le recueil jusqu'au déploiement, et ses perspectives d'évolution sont très encourageantes. Parmi les particularités de ce système, citons quelques points forts pour la gestion des connaissances sur des systèmes complexes. En premier lieu, Athanor offre trois originalités par rapport aux autres approches :

- gérer un triplet de connaissances : processus, système, compétences ;
- intégrer la phase ultime du déploiement ;
- intégrer une traduction automatique et transparente des modèles vers une base de connaissance opérationnelle, et en conséquence offrir une activité de résolution de problèmes grâce à des moteurs d'inférence.

D'autre part, la complexité des système à maintenir est prise en compte grâce : au couplage du modèle organique à des représentations en réalité virtuelle, à l'aide à la décision offerte par le modèle processus, et à la facilité de l'appropriation des modèles graphiques par les experts, et enfin à possibilité qui en résulte de réaliser une montée en charge progressive des connaissances stockées dans le serveur.

Enfin, d'un point de vue plus technique, l'architecture modulaire implémentée permet un déploiement progressif du serveur et facilite les extensions à de nouveaux modules de connaissances ; et les technologies de l'Internet utilisées simplifient le déploiement et la diffusion des connaissances et offrent une architecture ouverte permettant l'accès à des services externes comme des moteurs de recherche, la documentation technique numérisée, la nomenclature des composants, l'accès au stock ou à une référence d'achat…

L'approche ATHANOR a abouti au développement d'un système opérationnel de maintenance pour les systèmes complexes, dont une implantation a été réalisée à La Poste : SAMANTA (Système d'Aide à la MAiNtenance des Trieuses Automatiques – cf. [GUI 00] et [FOL 00]). Après s'être familiarisés avec l'éditeur de connaissances, les experts en machine de tri chargés de l'administration du serveur ont commencé à assurer la mise à jour et l'évolution des connaissances maintenues par l'outil. La phase de recueil des connaissances s'est étalée sur deux ans et s'est appuyée sur quatre experts. Elle a permis de mettre en évidence une trentaine de processus de diagnostic «haut niveau », nécessitant la construction d'un logigramme par processus. Les experts ont ainsi fait apparaître plus de 400 sommets tests et environ 200 diagnostics différents ont été répertoriés.

Références

[BAR 92] J.-P. Barthélemy and E. Mullet (1992). A Model of selection by aspects. *Acta Psychologica*, 79 :1-19, 1992.

[DIEN 99] R. Dieng, O. Corby, A. Giboin, and M. Ribiere (1999). Methods and Tools for Corporate Knowledge Management. *International Journal of Human-Computer Studies*, special issue on Knowledge Management, 51:567--598, 1999.

[ERM 96] J.-L. Ermine, M. Chaillot, P. Bigeon, B. Charreton, D. Malacieille (1996). MKSM : Méthode pour la gestion des connaissances. *Ingénierie des systèmes d'information*, vol. 4, pp 541-575, AFCET-Hermès, 1996.

[FOL 00] D. Follut, F. Guillet, P. Vandekerckhove and J. Philippe (2000). Samanta: Towards Using Virtual Reality in an Computer-Assisted Environment for the Maintenance of Postal Sorting Machines. In the *first French-British International Workshop on Virtual Reality*. July 11-12 2000, Brest, France.

[GUI 00] F. Guillet, D. Follut, P. Vandekerckhove, J. Philippé (2000),. Un serveur de connaissances dans un contexte de maintenance appliquée aux machines de tri postal. *Journées Internationales Ingénierie de systèmes et NTIC* (NimesTIC'2000), pages 30-35, Nîmes, 11-13 Septembre 2000.

[PEN 00] J.-M. Penalva (2000). Connaissances actionnables et intelligence collective. *Journées Internationales Ingénierie de systèmes et NTIC* (NimesTIC'2000), pages 2-11, Nîmes, 11-13 Septembre 2000.

[PHA 99] D.T. Pham, S.S.Dimov, R.M. Setchi (1999). Intelligent Product Manuals, *Proceedings of Institution of Mechanical Engineers*, Vol. 213, Part I, pp 65-76, 1999.

[SCH 94] A. Th. Schreiber, B. J. Wielinga, J. M. Akkermans, W. Van de Velde, and R. de Hoog (1994). CommonKADS: A comprehensive methodology for KBS development. *IEEE Expert*, 9(6), December 1994.

Annexe D

A System of Emotional Agent for Decision-Support[59]

In proceedings of : *IEEE/WIC/IAT Int. Conf. on Intelligent Agect technology (IAT'05)*, 2005. IEEE Press.

A system of emotional agents for decision-support

Stéphane Daviet[1,2], Hélène Desmier[1,2], Henri Briand[1], Fabrice Guillet[1], Vincent Philippé[2]
[1]LINA, [2]PerformanSe
stephane.daviet@knowesia.fr, helene.desmier@performanse.fr

Abstract

To facilitate the study of numerous phenomena, we use computer simulation tools. However, simulation of human behavior is still a challenge for both computer and human sciences. Multi-agent systems have already proved their efficiency to simulate complex interacting systems. In this context, we have worked on the simulation of human groups to study the emergence of behavioral patterns, mixing techniques of artificial intelligence, multi-agent systems and psychological sciences. This paper presents a new model of emotional agent, the Emotion, Feeling, Temperament agent (EFT). Based on BDI architecture, our model integrates the OCC emotional model and the PerformanSe behavioral model. We also present a concrete implementation of our model: the simulation of brain-damaged people's behavior on a production line. Then, we describe the modeling of the interactions of this system with AgentUML. The final goal of our work is to produce simulation data than can be enhanced through a KDD process.

1. Introduction

Usually in decision-support computing, tools are used to analyze the existing data of a company to extract knowledge [23]. In the "Tc&Plus.Virtuel" project, our objective is to build a decision-support tool for a company whose role is to accompany brain-damaged people in reintegrating the working world. Consequently, our goal is to get knowledge about these people working in a workshop and about their psychological evolution over a working day. Observing them *in vivo* would be very long and difficult because of the influence of the observer on the behavior of brain damaged people. A computer simulation, based on a multi-agent system [26], appeared to be relevant to get this knowledge by *in virtuo* experiments.

Based on Bratman's work [3], the BDI (Belief, Desire, Intention) architecture is the most used to build rational and cognitive agents. A BDI agent has a set of beliefs (information about its environment), a set of desires (what it wants) and a set of intentions (actions to do to reach its goal) [20]. Its control process follows this algorithm: 1) to perceive the environment, 2) to update beliefs and desires, 3) to choose an intention and 4) to act. Concerning emotions, since it is a recent subject of research, no consensus exists on a reference architecture as for the BDI model for rational agents. A first path of studies is the immersion of humans in a virtual world [5][25]. A second path, the one we have chosen to follow is the modeling of emotions for agents. Because of the nature of our application, we will focus on the second one. Several other projects investigate this path. The projects "Affective Reasoner" [15] and "PETEEI" [9] are interested in the social aspect of emotions without considering influence on reasoning. The "MRE" project is about emotions and planning but does not consider the notions of belief and desire. The BDE (Belief, Desire, Emotion) [12] agent is a BDI agent to which emotions are added. This method is the one we have chosen to build our new model of agent: the EFT (Emotion, Feeling, Temperament) agent.

This paper presents the work we realized to model and implement our prototype of simulation platform. We propose a model of emotional agents and of their interactions. This is the result of the collaboration between PerformanSe (skilled in behavioral tools), Aparta (implied in brain damaged people reintegration) and academic skills in computing and multi-agent systems. In the first part of this paper, we explain the context of the simulation and important notions of psychology, both general and specific to brain-damaged people, necessary to build a model. Then, we propose the model of our EFT agent based on the BDI model, emotional theories (OCC model [19]) and PerformanSe behavioral model. Then, we will present the modeling of the interactions of the system and their link with emotions thanks to the AgentUML formalism [18]. Finally, we will present the implementation of the model under the JADE platform [2] and the results we have obtained with this prototype.

2. Problem statement and preliminary concepts in psychology

2.1. Problem statement

PerformanSe already has 15 years of experience in designing tools for evaluating the behavior. It has also developed tools to determine *a priori* the dynamism of a group in term of attraction / repulsion forces. However, no

tools exists to simulate the evolution of the behavior of an agent and, even less so, of a group. Embedding the PerformanSe behavioral model into a virtual agent could be a way to obtain some results. We will see that the temperament is not sufficient to simulate the instant behavior of a person. Indeed, emotions have a first importance role in human reactions to environmental events.

2.2. Applicative framework

To construct a model of reusable agent adapted to any environment is quite an ambitious task. For the moment, to set the basis of our model, we have chosen to study a special case that will show the relevance of our simulation.

The context of the simulation is a production line situated in a protected workshop, Aparta. In this company, the employees suffer from brain-damage after an accident and are here to learn how to reintegrate society.

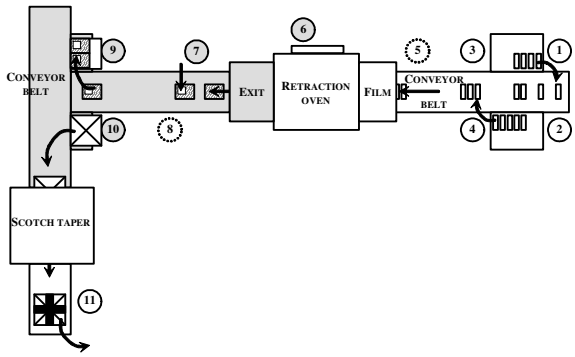


Figure 1. Packaging production line.

The modeled system is production line to repackage packs of cookies together (Figure 1). On this line, employees 1 and 4 put packs of cookies on the conveyor belt while 2 and 3 are supplying packs to 1 and 4. Then, these packs are wrapped in a film and go through an oven to be packaged in fours. Employee 6 is a non brain-damaged person and is responsible for the good functioning of the line. Moreover, he is there to encourage tired or unmotivated people. Then, at the end of the oven, employee 7 puts a label on the package. Then, the packages arrive in front of employees 9 and 10 who put them into cartons for the transport. Finally, employee 11 closes the cartons and prepares them to leave the manufacture.

For our first implementation, we chose a subset of this line (the grey part) composed of employees 6, 7, 9 and 10. The choice is based on the variety of interactions contained by this subset. Indeed, we encounter both human-to-human and huma-to-machine interactions.

Aparta is a particular company. It has the same kind of constraints of performance or efficiency as any firm but it also has to succeed in its reintegration mission. Discovering less stressing workstations on the line or groups with people getting on well together, for instance, is a major stake. It explains the interest of our simulation.

2.3. Specificities of brain-damaged people and emotional stake

The originality of this simulation is its application in an environment of brain-damaged workers. Therefore, we had to understand the effects of the most common injuries to the brain. The main specificity of a brain-damage lesion is that emotional filters disappear or are less efficient [16]. Moreover, brain-damaged people suffer from lapses of memory and concentration. Then, recent discoveries in neuropsychological studies [7] pointed out the necessity of an emotional process to be able to think and decide. This justifies the need we had to build convincing virtual agents with emotions.

The first step of validation is to succeed in the realization of spontaneous and unpredictable behaviors specific to brain-damaged people, that are more reactive, because their emotional filters have disappeared, than those of non brain-damaged person.

2.4. Existing behavioral models

The human behavior is strongly influenced by emotions and the general psychological state of the person. It is the origin of non-rational reactions to an event, of individual specific behavior and partly of social interactions within a group. Therefore, to simulate human behavior, we have to consider two complementary aspects: emotion and personality.

We can distinguish four main types of emotional theories [6]: Darwinist, Jamesonian, cognitivist (appraisal theory) and social constructivist. The cognitivist approach links the triggering of emotions with how the person appraises its environment.

One of the most widespread cognitivist models of emotions is that of Ortony, Clore and Collins, often named OCC model [19]. They have sorted 22 emotions (11 pairs (Table 1)) in three main classes: those linked with the relevance of the consequences of an event compared to the expectations of the agent (most often its goals), those connected with the moral value of the actions of another, and those linked with the affection of an object. The model is especially designed for computer use. Therefore, most of the projects about emotional agents use this model with a few minor modifications [15][9][21]. In our case, this model is especially appropriate because the system is closed and the number of external events perceived by an

agent is limited. Determining the emotional consequences of each event is then easier than in an open system where unpredictable external events could happen.

Table 1. The eleven pairs of the OCC model.

Happy-for	Resentment
Gloating	Pity
Joy	Distress
Pride	Shame
Admiration	Reproach
Love	Hate
Hope	Fear
Satisfaction	Fear-confirmed
Relief	Disappointment
Gratification	Remorse
Gratitude	Anger

For the behavioral aspect, we have worked on the model of the PerformanSe company [13]. It is based on the systemic theory developed in the fifties by the School of Palo Alto [11]. The PerformanSe model has followed this theory considering the person as a system interacting with a context (environment, family and working circles, working rules, etc.) [13]. This system is defined by ten behavioral dimensions (Table 2) interacting with each other and with the context presented.

This model is an extension of the "Big Five" [11] model which has concluded that personality traits are linked with the behavior. Three motivations complete this model and reflect the major trends a person would follow in his/her behavior. The PerformanSe model has proven its validity with multiple studies on representative panels.

Table 2. The ten behavioral dimensions.

Extroversion	Introversion
Anxiety	Relaxation
Assertiveness	Questioning
Receptiveness	Determination
Rigor	Improvisation
Intellectual dynamism	Intellectual conformism
Combativeness	Conciliation
Motivation of achievement	Motivation of facilitation
Motivation of belonging	Motivation of independence
Motivation of power	Motivation of protection

3. Modeling of the EFT agent

3.1. The need of introducing feelings

Our emotional agent is built on the BDI model (Figure 2). Its personal features stand for the set of inner behavioral characteristics of the agent. First, the originality of our model is the agent personality divided into three levels: emotions, feelings and temperament. We based this separation on the variability of each level. Emotions are ephemeral, representing instant feelings. For instance, you only feel frightened for a relatively short time. Concerning temperament, it represents the nature of our agent. It is supposed to be constant in time. Indeed, if we consider "extroversion", we assume this feature does not evolve. Just keeping these two parts did not enable our agent to remember its interactions or what it felt. For example, an agent is more likely to become friendly with an agent that gives him happiness that with another. That is why we introduced the notion of the feelings (friendship, for example). Feelings evolve with the emotions felt and in function of the agent's temperament.

3.2. Description of the EFT part of our model

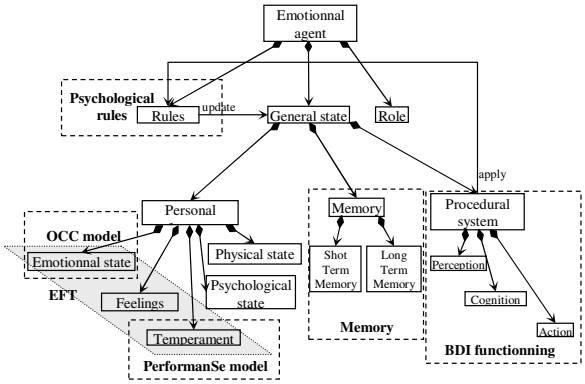


Figure 2. UML class diagram of the EFT agent.

To build the personality of an agent, we used proven models. Emotions are based on the OCC model as described above with its 11 pairs of emotions (Table 1). We complete this model by conditioning the generation and the intensity of an emotion by the temperament. For instance, somebody who is rather anxious and extroverted will be angry more easily than an introverted and relaxed one. To model the temperament, PerformanSe behavioral model is used. Since the temperament does not evolve in our simulation, it will always be used as a conditional parameter of psychological rules. Relating to feelings, we did not find any model.

Therefore, we worked with psychologists to develop a model of six pairs of feelings presented in Table 3. This notion is interesting because it enables each agent to have its own personality (personal feelings and emotions) but it can also have several types of relationships with the other agents (relative feelings and emotions). Emotions and feelings relative to someone else are part of the social knowledge of an agent. Indeed, it corresponds to a part of the beliefs of the agent that will determine its social relations. The difference with usual beliefs of the BDI

model is that those are not the result of a process of the agent to apprehend its environment. They are only the result of its social interaction with other agents. Another point is that feelings and emotions affect the way the agent chooses its goals and intentions in a non-rational way contrary to the BDI model.

Table 3. Six pairs of feelings for the EFT agent.

Quietude	Worry
Self respect	Low regard
Friendship	Hostility
Trust	Distrust
Indebtedness	Rancor
Compassion	Indifference

3.3. Control process of the agent

The procedural system corresponds to the cognitive functioning of the agent and is based on BDI architecture. First, the agent perceives its environment. The perception is filtered according to the personal state (a relaxed agent will be less sensitive to malicious remarks) and physical state (an agent with a little deafness will less perceive a loud noise). Then, the agent updates its beliefs in consequence of the filtered message, decides what to do and then acts.

Finally, a base of psychological rules (established with experts) links all these parts. These rules manage personality evolution and influence, choice of actions and kind of communications. For instance, rules can be expressed in the following manner: "If I am very stressed and very anxious and angry and my neighbor is speaking to me harshly, therefore I will shout at him".

The control process of the agent is its functioning. It enables the agent to act, talk or react. Talking and acting are parallel processes. In our case, communication is not a means for the agent to collaborate or to negotiate with other agents. We only use communication as a vector to have our agents exchange their feelings. In this way, we cannot really consider exchanged messages as speech act [1][24], they are not a mean for the agent to achieve its goals. It is the reason why we have separated the two processes. This choice simplifies protocol management and thus, an agent can act and speak at the same time, which increases its reactivity. In this part, we will just describe the control of acting (Figure 3) since speaking is a very simple process (get the message, update the personal state and generate possible answers).

First, all the perceptions (messages) of an agent are stocked in a buffer and analyzed one by one. The first message of the buffer is analyzed (filter and update of beliefs, desires and internal state). Then, the strongest desire becomes the intention of the agent. At this moment, it checks if its current goal is still in adequation with its intention. These two steps are important because they enable an agent to change its intention before completing it (the change is between two elementary actions of the plan). We talk about a BDI agent with open-minded commitment [20] – *i.e.* it will maintain an intention as long as it is still believed possible. Thus, the agent is very reactive to the environment evolution. After that, if the goal corresponds to the intention, the agent acts. Otherwise, it changes its goal to fit its intention.

We can see in Figure 3 that in plan search the agent has the ability to construct individual plans and collaborative plans to reach its goal. We do not yet work on this part which implies planning and requires a lot of investigations to have emotions impact on it.

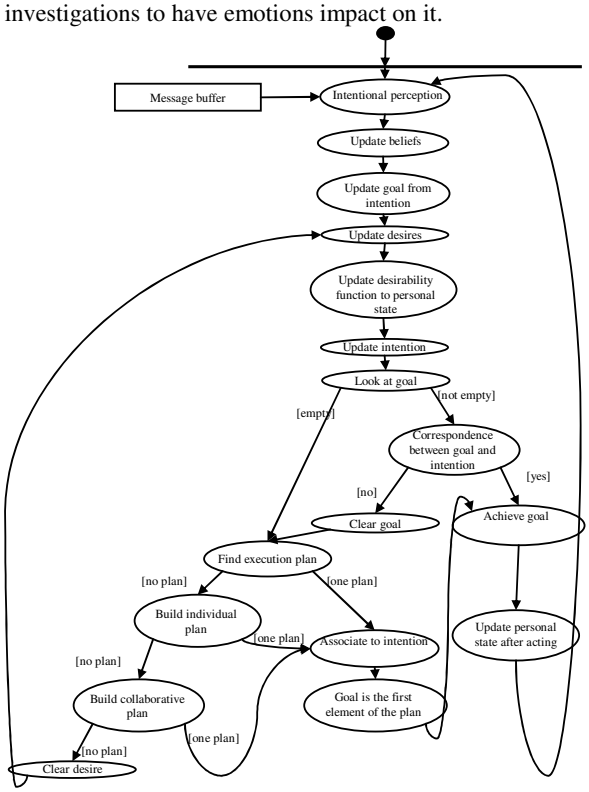


Figure 3. Control process of the EFT agent.

We will now insist on how emotions and more generally how internal state take place in this cognitive architecture. The main influence is on reasoning. Indeed, emotional variables affect the action chosen by the agent. Following its personal state, an agent will not have the same choice of possible actions. On the other hand, relationships are memorized thanks to feelings. Therefore communication is also influenced by the internal state of an agent. Temperament is necessary to enable each agent to react differently to the same event. Finally, emotional state is strongly linked to the physical state which enhances the possible relational situations.

4. Modeling of the system using Agent UML

The work around AgentUML began with the difficulties to model multi-agents systems only with UML [18]. AgentUML consists in adapting UML diagrams to oriented agent programming. The main propositions concern interactions diagrams. The conditional structures are simpler and modeling communication between agents is easier. Moreover, AgentUML introduces the notion of interaction protocols which give a solution to formalize exchanged messages and are reusable. That is why, after using UML for the agent, we have used AgentUML to model the system.

4.1. Interactions of the system

The interaction aspect of an MAS is crucial. Indeed, *"there's no such a thing as a single multi-agent system"* [26]. Interactions and agent communication are the basis of the emergence of group behavior, particularly interesting in our case. What we are interested in is seeing the evolution of someone interacting with the machine and with its co-workers. Thus, it becomes necessary to describe the set of primary interactions that leads a person to evolve. To achieve this goal, we have both worked with psychologists (expert of brain-damaged people) and observed the behavior of the workers on the line.

We have followed a three-step top-down methodology. First, we have designed a global overview of the system with UML use-case diagrams (Figure 4). The second step consists in describing the activity of the system with AgentUML activities diagrams [18]. Finally, we have detailed each of the previous activities diagrams with AgentUML sequence diagram.

We have chosen a number of key interactions of the system, which are characteristic of the modeled environment and source of the dynamism of social relations. Our simulation does not focus on language; we privilege the reactive aspect and its impact on emotions, to the dialog. We have distinguished four classes of interactions: to communicate with somebody, to appear to somebody, to act physically on the environment, and to perceive its environment.

Emotions influence the way people talk to each other both in the wording and in the manner the sentence is expressed – tone, velocity, loudness. Moreover, a remark could be positive (compliment) or negative (reproach). So we have introduced an intention coefficient – malevolence/benevolence of message and distinguished these two types of remark.

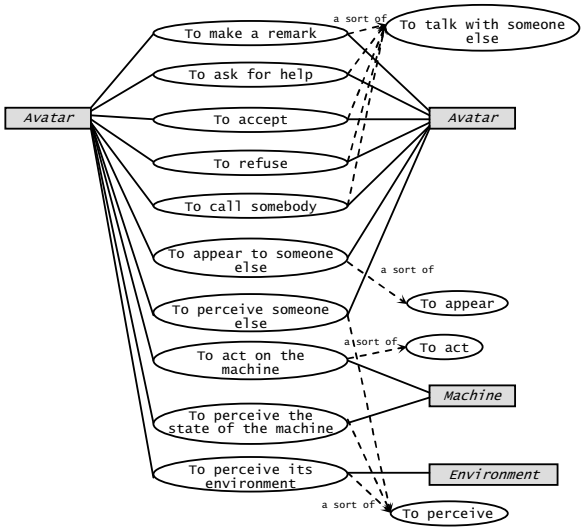


Figure 4. General UML use-case diagram.

4.2. Protocols of interaction

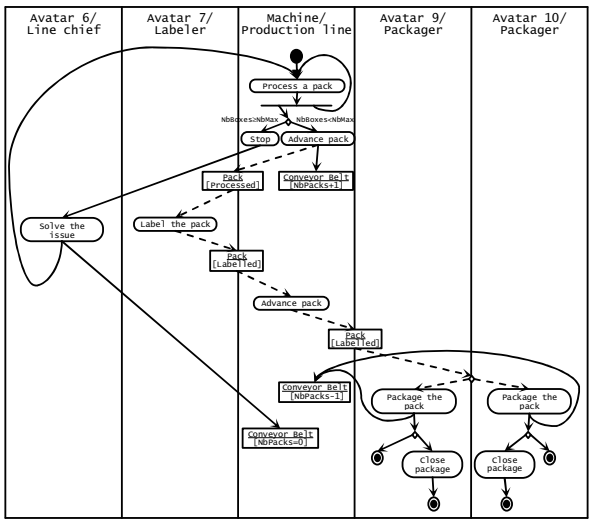


Figure 6. General UML activities diagram.

We have first designed an activity diagram to describe the general production line functioning (Figure 6), and specific activity diagrams for each workstation of the line (workstation 7 in these examples).

We have then designed general protocols shared by all the human agents of the system. The "voluntary perception" protocol describes the fact that an agent decides to perceive another agent (human or machine). While it is true that in an environment, you also receive unsolicited stimuli (like noise, light). It is particularly – important in our case to make this distinction because emotional state of brain-damaged people is strongly affected by unsolicited stimuli due to their weak capacity

of concentration. Moreover, a workshop conveys a lot of perturbation (noise of the machine, excessive or insufficient light for instance). The general "working protocol" describes how an agent interacts on the production line. For each workstation of the line, we have realized a sequence diagram of the working protocol describing precisely sequence of the tasks to accomplish (Figure 7). The last and richest type of interaction is the communication between people. We can distinguish "remark", "help request" and "inform" protocols.

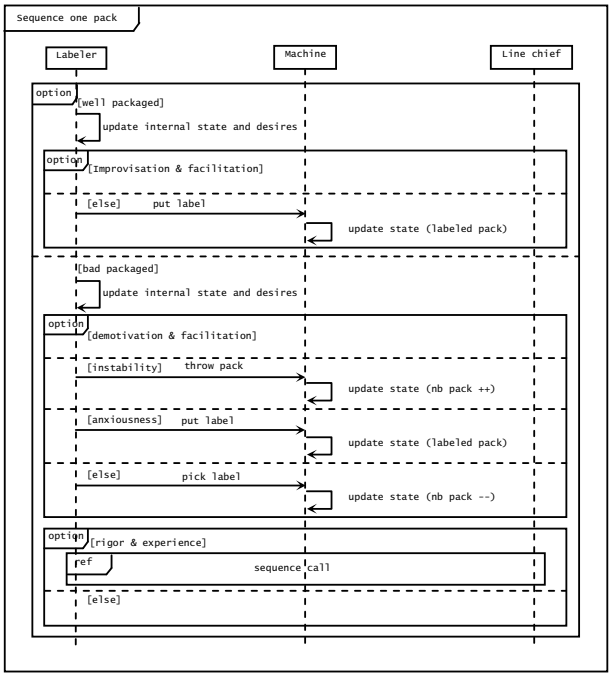


Figure 7. AgentUML sequence diagram for workstation7 (sequence one pack).

5. Realization

5.1. The prototype of the implemented simulation

At the first stage of this project, only a part of the model proposed was implemented under the widely used JADE platform [2]. Three classes of comportment have been defined: paranoid, well-adapted and ill-adapted. Each one is described with the three personality traits: extroversion, anxiety and combativeness. It contributes to the validation of the prototype because these behaviors are quite easy to detect on a simulation by looking at the evolution graphs of emotional parameters. Reproducing these behaviors partly proves the value of the model and of the rules.

We present on Figure 8 screenshots of the simulation. An agent is represented by an expressive "Smiley". Illustrative speech balloons represent messages. Each agent has a window to represent its internal state: temperament, emotions and feelings.

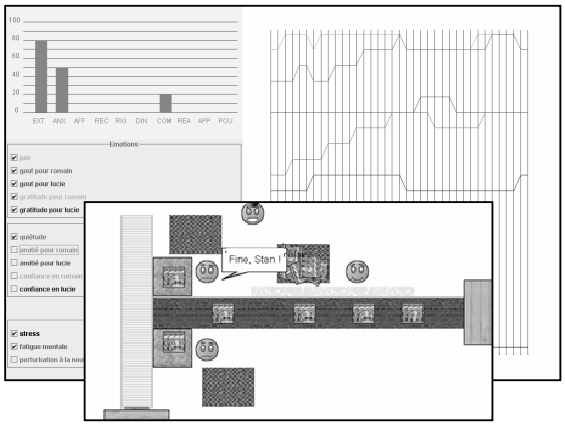


Figure 8. Screenshot of the main screen.

5.2. Results and expectations

This application gives promising results. We have reached our first goal that was to observe different behaviors for different initializations of the agents. For instance, a much-stressed paranoid agent will throw unlabelled packs whereas a conscientious agent will put them away. Moreover, on the larger scale of the overall system, we have observed there were differences both in terms of work efficiency and psychological "well-being" depending on the composition of the group of agents. The simulation has already led to modifications on the line to favor the rotation of workers at a much-stressing workplace.

Now, the first thing to envision is to get much more psychological expert evaluation and then to add new possible interactions between agents and with the environment. The main difficulty here is that the number of emotional rules to collect increases with the number of the possible actions of the agent. Indeed, each action must have its own triggering rules. At this point, we could envision two solutions: determining some meta-rules for meta-classes of actions or thinking to a less determinist system that does not need this kind of rules. In all cases, it is essential to make a simulation more exhaustive and consequently more reliable for the deciders. We will have to work with them to evaluate and satisfy their new requirements that will appear along with an improvement of the simulation.

6. Conclusion

In this paper, with the idea to build a decision-support application to encourage the reintegration of brain-

damaged people, we have proposed an emotional agent model. The notion of emotion was developed along three axes: emotions, feelings and temperament (EFT agent). The distinction of these three axes based on the duration of the evolution of each dimension is the keystone of our model. More important than the emotions, feelings and temperament parameters proposed, the idea of splitting the emotional aspect into multiple dimensions can be a good basis to develop convincing virtual humans. Then, in our model, the psychological aspect plays a role at different levels of the control process: cognitive, perceptive and expressive. It is an important stake to be able to simulate all the aspects of emotions on the behavior.

In the short term, our objective is to complete psychological appraisal to enhance the simulation and, consequently, the complexity of potential behaviors. Then, we will have to study emergent behaviors and compare them to real-life behaviors already observed. Then, it will be interesting to extend the study to non-brain-damaged people. We have built a model that can take into account the specifities of brain-damaged people, but that is also sufficiently generic to start applying the same kind of approach to non brain-damaged people. In a second time, it would also be interesting to adopt a methodology to evaluate our emotional model as done in [14].

7. References

[1] J. L. Austin, How to do things with words, Oxford University Press, 1962.

[2] F. Bellifemine, A. Poggi, G. Rimassa, JADE-A FIPA-compliant agent framework, in *Proc. Of the PAAM 99*, 1999, pp. 97-108.

[3] M. Bratman, *Intentions, Plans and Practical Reason,* Harvard University Press, 1987.

[4] L. Braubach, A. Pokahr, D. Moldt, W. Lamersdorf, Goal representation for BDI agent systems, in *The second international workshop on Programming Multiagent Systems (PROMAS 04)*, 2004.

[5] A. Camurri, A. Coglio, An architecture for emotional agents, *IEEE Multimedia*, 1998, 5(4):24-23.

[6] R. R. Cornelius. Theoritical approaches to emotion, in *SpeechEmotion-2000*, 2000, pp. 3-10.

[7] A. Damasio*, Descartes' error: emotion, reason and the human brain*, Putnam Books, 1994.

[8] P. Ekman, W. Friesen, P. Ellsworth, *Emotion in the human face*, Cambridge University Press, 1972.

[9] M. S. El-Nasr, T. R. Ioerger, J. Yen, A web of emotions, in *Proc. Of the workshop on Emotion-Based Agent Architecture (EBAA 99) at the 3rd International Conference On Autonomous Agents (Agents 99),* 2000.

[10] FIPA, FIPA ACL Message Structure Specification, FIPA Specifications, http://www.fipa.org/specs, 2000.

[11] D. W. Fiske, Consistency of the factorial structures of personality ratings from different sources, *Journal of abnormal and social psychology*, 1949, 44:329-344.

[12] A. M. Florea, E. Kalisz, Behavior anticipation based on beliefs, desires and emotions, in *Proc. Of the 6th international conference on Computing Anticipatory Systems (CASYS 03)*, 2003.

[13] R. Gras, P. Peter, S. Baquedano, J. Philippé, Structuration de comportements de réponse à un questionnaire par des méthodes multi-dimensionnelles, *Extraction et gestion des connaissances, extraction des connaissances et apprentissage*, 2003, 17(1-3) :105-118.

[14] J. Gratch, S. Marsella. Evaluating the modeling and use of emotion in virtual humans, in *Proc. Of Autonomous Agents and Multi-Agent Systems (AAMAS O4)*, 2004.

[15] A. Kapoor, S. Mota, R. W. Picard, Towards a learning companion that recognizes affect, in *Proc. Of emotional and intelligent : the tangled knot of social cognition, AAAI Fall symposium*, 2001.

[16] M.V. Der Linden, X. Seron, DL Gall, P Andrès, *Neuropsychologie des lobes frontaux*, Solal, 1999.

[17] J. Mélèze, *L'analyse modulaire des systèmes de gestion, A.M.S*, Hommes et techniques Ed., 1972.

[18] J. Odell, H. Van Dyke Parunak, B. Bauer, Extending UML for agents, in *Proc. Of the agent-oriented information systems workshop at the 17th National Conference on Artificial Intelligence (NCAI 00)*, 2000, pp. 3-17.

[19] A. Orthony, G. Clore, A. Collins, *The cognitive structure of emotions*, Cambridge University Press, 1988.

[20] A. Rao, M. Georgeff, Modeling rational agents within a BDI-architecture, in *Proc. Of the 2nd International Conference on Principles of Knowledge Representation and Reasoning (KR 91)*, Morgan Kaufmann Publishers Inc, 1991, pp. 473-484.

[21] W. S. Reilly, *Believable social and emotional agents*, PhD Thesis, Carnegie Mellon university, Pittsburgh, 1996.

[22] J. Rickel, S. Marsella, J. Gratch, R. Hill, D. Traum, W. Swartout, Steve goes to Bosnia : towards a new generation of virtual humans for interactive experiences, *IEEE Intelligent Systems*, 2002, 17(4):32-38.

[23] V. Sandoval, *L'informatique décisionnelle*, Hermès Sciences, 1997.

[24] J. R. Searle, Speech acts: an essay in philosophy of language, Cambridge University Press, 1969.

[25] J. Velasquez, Modeling emotions and other motivations in synthetic agents, in *Proc. Of the 4th International Workshop on Agent Theories, Architecture and Languages (ATAL 97)*, 1997.

[26] M. Wooldridge, *An introduction to multi-agents systems*, Wiley, 2001.

Annexe E

Conceptual hierarchies matching : an approach based on discovery of implication rules between concepts [54]

In proceedings of : *17th European Conference on Artificial Intelligence (ECAI)*, IOS Press, 2006.

Conceptual hierarchies matching : an approach based on discovery of implication rules between concepts

Jérôme David and **Fabrice Guillet** and **Régis Gras** and **Henri Briand**[1]

Abstract. Most research works about ontology or schema matching are based on symmetric similarity measures. By transposing the association rules paradigm, we propose to use asymmetric measures in order to enhance matching. We suggest an extensional and asymmetric matching method based on the discovery of significant implications between concepts described in textual documents. We use a probabilistic model of deviation from independence, named implication intensity. Our method is divided into two consecutive stages: (1) the extraction in documents of relevant terms for each concept; (2) the discovery of significant implications between the concepts. Our method is tested on two benchmarks. The results show that some relevant relations, ignored by a similarity-based matching, can be found thanks to our approach.
Keywords: ontology matching, term extraction, implication intensity, extensional matching, association rules.

1 Introduction

With the increase of electronic data and knowledge on the Internet or in companies, the hierarchical categorization of data through ontological forms as taxonomies has been widely used. Web directories such as Yahoo.com and OpenDirectory, the Electronic Document Management, or the Semantic Web with its OWL ontology are examples of such taxonomies.

In the literature, a lot of works deals with schema/ontology matching. The schema or ontology matching aims at finding semantic relations (i.e. equivalence, subsomption, etc) between entities (i.e. concepts, properties) of two schemas/ontologies. These approaches use various techniques such as machine learning [7], FCA-Analysis [17], database schema matching [12], graph matching [13]. These approaches are commonly based on similarity measures for discovering equivalence relations between concepts.

However, the extracted matchings can be enhanced by using asymmetric measures, which deliver more accurate information in the form of implications between concepts. In knowledge discovery in databases (KDD), asymmetric measures, called interestingness measures, are widely used for association rules discovery [1]. Association rules are expressions of the type "if *antecedant* then *consequent*" representing implicative tendencies between conjunctions of attributes in databases. In this paper, we evaluate the use of such asymmetric measures for matching concepts of schemas or ontologies by using the Implication Intensity [9, 2], a probabilistic model of deviation from statistical independence.

Our matching method is both extensional and terminological. It is designed to be used on taxonomies of concepts associated with textual documents. The idea underlying our approach considers that one concept is more specific than another, if the vocabulary used in the documents associated to the first concept tends to be included in the vocabulary of the other one.

Our method is divided into two consecutive stages: (1) The extraction of concept-relevant terms; (2) The discovery of association rules between concepts.

This paper is organized as follows. In a first section we give an overview of matching approaches. Then, we introduce the Implication Intensity measure, and the concept hierarchy model, and the used formalism. Next, we apply the two stages of our method focusing on rule extraction. Finally, we experiment our method on two datasets and discuss the results obtained.

[1] LINA CNRS FRE 2729, Polytechnic School of Nantes University, France, email: jerome.david@polytech.univ-nantes.fr

2 Related works

Many surveys about ontology and schema matching have been proposed in litterature [10], [15], [16]. The two last ones propose a classification and a comparative study of matching approaches. The survey [15] focuses on the database schema matching approaches, while [16] reuses this classification for ontology matching. From these surveys we can distinguish: the extensional approaches (or element-based), and the intensional approaches (or only-schema-based). The matching approaches can be also discriminated regarding the kind of relations that they are based on. Some consider symmetric (equivalence) relations, while other ones also use asymmetric relations such as the subsomption or implication.

The main part of these works propose to process the concept name by using string-similarities (Anchor-PROMPT [14], Cpuid [12], Coma [5], S-MATCH [8]) or/and external oracles such as Wordnet ([8]). They can also use the schema or ontology structure (Similarity Flooding [13], Artemis [3], [5], [14], [12]).

Most of these approaches are intensional and symmetric. None of them are both asymetric and extensional. Among extensional approaches, we can cite GLUE [6]. This symmetric approach uses Bayesien learners in order to classify instances of the first ontology into the other and vice-versa in order to estimate the joint probability distribution and then predict concept similarities.

We can also notice that there is only one intensional method considering asymmetric relations. The method S-MATCH [8] searches equivalence ($=$) relation between concepts but also the more general ($\sqsupseteq$), less general ($\sqsubseteq$), mistmatch ($\perp$) and overlapping ($\sqcap$) relations. This method uses a lot of single matchers: 13 linguistic-based matchers and 3 logic-based matchers.

3 The definition of the Implication Intensity

Let us now consider a finite set T of n individuals described by a set I of p items. Each transaction t can be considered as an itemset so that $t \subseteq I$. $A = \{t \in T; a \subseteq t\}$ is the extension of itemset a and $\overline{B} = T - \{t' \in T; b \subseteq t'\}$ is the extension of $\overline{b}$. Then, we introduce the quantities $n_a = card(A)$, $n_{\overline{b}} = card(\overline{B})$ and $n_{a\wedge\overline{b}} = card(A \cap \overline{B})$.

An association rule [1] is an implication of the form $a \rightarrow b$, where a and b are disjoined itemsets. In practice, it is quite common to observe a few transactions which contain a and not b without having the general trend to have b when a is present contested. Therefore, the number $n_{a\wedge\overline{b}}$ of counter-examples must be taken into account to statistically accept to retain or not the rule $a \rightarrow b$.

More precisely, we compare the observed number of counter-examples $n_{a\wedge\overline{b}}$ to a probabilistic model noted $N_{a\wedge\overline{b}}$. Let us assume that we randomly draw two subsets X and Y in T which respectively contain n_a and n_b transactions, i.e. $N_{a\wedge\overline{b}} = card(X \cup \overline{Y})$.

The implication intensity of the association rule $a \rightarrow b$ is defined by:

$$\varphi(a \rightarrow b) = 1 - \Pr(N_{a\wedge\overline{b}} \leq n_{a\wedge\overline{b}}) \quad (1)$$

The distribution of the random variable $N_{a\wedge\overline{b}}$ depends on the drawing mode [9]. It is establish that, under some conditions, the random variable $N_{a\wedge\overline{b}}$ follows a Poisson law with $\lambda = n_a n_{\overline{b}}/n$.

4 Concept hierarchy model and formalisms used

Our approach (figure 1) is designed for conceptual hierarchies of concepts organized by a partial order relation, connected to a set of textual documents.

We define a conceptual hierarchy $\mathcal{H}$ as a quadruplet:

$$\mathcal{H} = (C, \leq, D, \sigma_0) \quad (2)$$

where c is a set of concepts, $\leq$ represents the partial order, D is the set of documents, and σ_0 is the relation which associates a set of documents to each concept (i.e. for a concept $c \in C$, $\sigma_0(c)$ represents the documents associated to c). From the partial order $\leq$, we extend the relation σ_0 to σ, where:

$$\sigma(c) = \bigcup_{c' \leq c} \sigma_0(c') \quad (3)$$

Then, we transform the hierarchy $\mathcal{H}$ defined on documents in a hierarchy $\mathcal{H}'$ defined on terms as follows:

$$\mathcal{H}' = (C, \leq, T, \gamma_0) \quad (4)$$

where T is the set of relevant terms extracted from D, and $\gamma_0 \subseteq C \times T$ is the relation associating terms to concepts (i.e. $\gamma_0(c)$ represents the set of relevant terms selected for the concept c). Technically, γ_0 is deduced from σ and the relation δ linking terms to documents (i.e. $\delta(t)$ is the set of documents in which the term t appears). From the partial order $\leq$, we extend the relation γ_0 to γ, where:

$$\gamma(c) = \bigcup_{c' \leq c} \gamma_0(c') \quad (5)$$

The common term set of two hierarchies $\mathcal{H}'_1 = (C_1, \leq_1, T_1, \gamma_1)$ and $\mathcal{H}'_2 = (C_2, \leq_2, T_2, \gamma_2)$ is noted $T_{1\cap2} = T_1 \cap T_2$. Next, we define the relation $\gamma_{1\cap2}$ which associates a subset of $T_{1\cap2}$ for each concept $c \in C_1 \cup C_2$:

$$\gamma_{1\cap2}(c) = \begin{cases} \gamma_1(c) \cap T_2 \; if \; c \in C_1 \\ \gamma_2(c) \cap T_1 \; if \; c \in C_2 \end{cases} \quad (6)$$

An implicative match set between two hierarchies $\mathcal{H}'_1$ and $\mathcal{H}'_2$ is a set of implicative rules. A rule $a \rightarrow b$ between the concepts $a \in C_1$ and $b \in C_2$ represents a quasi-implication from the set of terms $\gamma_{1\cap2}(a)$ to the set of terms $\gamma_{1\cap2}(b)$.

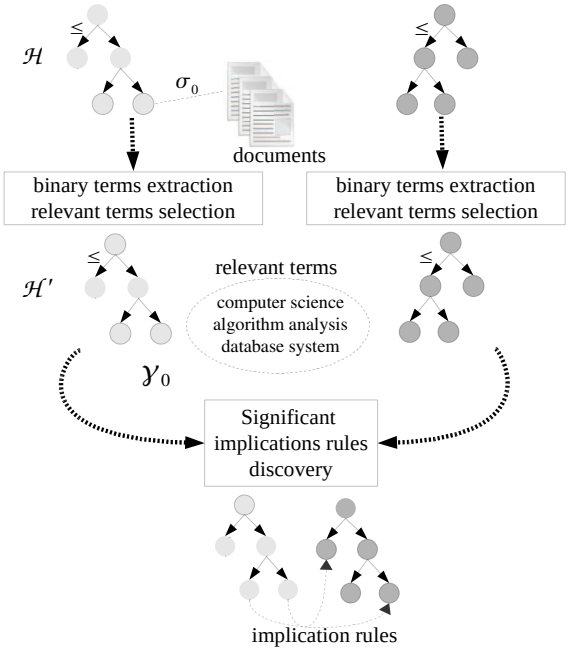


Figure 1. Methodology scheme

5 Extraction and selection of relevant terms

The goal of this process is to extract terms from the documents and then select relevant terms for each concept of the two hierarchies. A term t will be relevant for concept c if t tends to appear in the documents associated with concept c. We choose to associate the term t with concept c if the rule $t \rightarrow c$ has an implication intensity value greater than a given threshold φ_t.

In order to evaluate the rules $t \rightarrow c$, we first extract T_0, the set of the binary terms (terms composed of two meaningful words) and of the verbs contained in the documents. Binary terms have the advantage to be less ambiguous than simple words. The acquisition of binary terms is performed by software program ACABIT [4]. The textual data are firstly POS-tagged and stemmed by the the software program Montylingua [11]. The figure 2 summarizes the term extraction process.

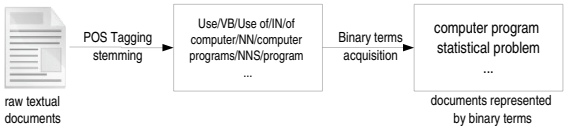


Figure 2. Terms extraction process

Next, we compute all the rules $t \rightarrow c$ and the selection of the relevant terms set of a concept c (noted $\gamma_0(c)$) is defined as follows:

$$\gamma_0(c) = \{t \in T_0 | \varphi(t \rightarrow c) \geq \varphi_t\} \quad (7)$$

where φ_t is the implication intensity threshold value and $\varphi(t \rightarrow c)$ is the implication intensity values of the rule $t \rightarrow c$ defined by:

$$\varphi(t \rightarrow c) = 1 - Pr\left(N_{t \wedge \bar{c}} \leq n_{t \wedge \bar{c}}\right) \quad (8)$$

$n_{t \wedge \bar{c}} = card(\delta(t) - \sigma(c))$ is the observed number of counter-examples, that is to say documents which contain the term t and which are not associated with the concept c. And $N_{t \wedge \bar{c}}$ is the expected number of counter-examples under independence hypothesis.

6 Discovery of significant rules between concepts

6.1 Selection criteria of significant rules

In section 4, we have defined a match result as a set of implication rules between concepts issued from two hierarchies $\mathcal{H}_1$ and $\mathcal{H}_2$. Nevertheless, a lot of rules can be discovered. In this section, we define the implication intensity of a rule between concepts, and then we give two criteria defining the notion of significant rule.

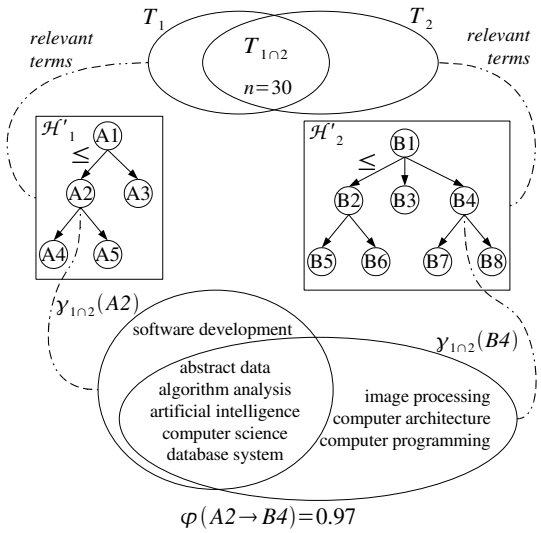


Figure 3. Evaluation of significant rules

The implication intensity of a rule $a \rightarrow b$ (with $a \in C_1$ and $b \in C_2$) is defined by:

$$\varphi(a \rightarrow b) = 1 - Pr\left(N_{a \wedge \bar{b}} \leq n_{a \wedge \bar{b}}\right) \quad (9)$$

where $n_{a \wedge \bar{b}} = card(\gamma_{1 \cap 2}(a) - \gamma_{1 \cap 2}(b))$ is the number of relevant terms for concept a which are not relevant for concept b. $N_{a \wedge \bar{b}}$ is the expected number of relevant terms for concept a which are not relevant for concept b. On figure 3, the rule $A2 \rightarrow B4$ has $n_{A2 \wedge \overline{B4}} = 1$ counter-examples. Its implication intensity value is:

$$\varphi(A2 \rightarrow B4) = \sum_{k=0}^{n_{A2 \wedge \overline{B4}}} e^{-\lambda} . \frac{\lambda^k}{k!} = 0,97$$

where $\lambda = n_{A2}.n_{\overline{B4}}/n = 6 \times 22/30$ (see figure 3).

Thus, the two criteria defining a significant rule are, first, its implication intensity value and, second, the specificity of its consequent combined with the generality of its antecedent. A rule $a \rightarrow b$ (with $a \in C_1$ and $b \in C_2$) will be significant if:

$$\varphi(a \rightarrow b) \geq \varphi_r \quad (10)$$

$$and\ \forall x \geq a,\ \forall y \leq b,\ \varphi(x \rightarrow y) \leq \varphi(a \rightarrow b) \quad (11)$$

The second criterion (equation 11) selects only generative rules and then permits to reduce redundancy in the extracted rules set. Indeed, from a significant rule $a \rightarrow b$, we can deduce all the rules of the form $u \rightarrow v$ (with $u \leq_1 a$ and $b \leq_2 v$) because at the term level: $\gamma_{1 \cap 2}(b) \subseteq \gamma_{1 \cap 2}(v)$ and $\gamma_{1 \cap 2}(u) \subseteq \gamma_{1 \cap 2}(a)$. We say that the rule $a \rightarrow b$ is generative of the rules set $u \rightarrow v$. For example (figure 3), the rule $A2 \rightarrow B4$ is generative of the rules set $\{A2 \rightarrow B1, A4 \rightarrow B4, A5 \rightarrow B4, A4 \rightarrow B1, A5 \rightarrow B1\}$.

6.2 Algorithms for rule extraction

During the rule extraction step, we can reduce the computation time with the help of the partial order. A top-down search phase enables us to avoid the evaluation of rules having too specific antecedents. This section presents our selection strategy divided into two algorithms.

Inputs :
  a : a concept of $\mathcal{H}_1$.
Input/Output variables :
  $\mathcal{B}_{current}$: a set of concepts taking from $\mathcal{H}_2$.
  $ruleList$: the list of selected rules.
Procedure specializeAntecedent(a, $\mathcal{B}_{current}$, $ruleList$)
Begin
  ForEach $b_x \in \mathcal{B}_{current}$ **Do**
    $specializeConsequent(a, b_x, \mathcal{B}_{current}, 0.0, ruleList)$
  End Do
  ForEach $child \in children(a)$ **do**
    $\mathcal{B}'_{current} := \mathcal{B}_{current}$
    $specializeAntecedent(child, \mathcal{B}'_{current}, ruleList)$
  End Do
End

Figure 4. Algorithm specializing the antecedent

Our first algorithm (figure 4) takes in a concept a from the hierarchy $\mathcal{H}_1$ and a set of concepts $\mathcal{B}_{current} \subset C_2$ from $\mathcal{H}_2$. For each concept of $\mathcal{B}_{current}$, the second algorithm (figure 5) searches and selects valid consequents. It also updates the set $\mathcal{B}_{current}$. And then, this first procedure is recursively launched over the children of a and with a copy of the set $\mathcal{B}_{current}$. The set $\mathcal{B}_{current}$ contains the subtrees of $\mathcal{H}_2$ with concepts that were selected during the previous recursion steps.

The second algorithm (figure 5) searches a set of valid consequents for the current antecedent a. The search is performed over the set candidate consequents $\{B_x | B_x \leq_2 B\}$. A consequent b_s will be selected if the rule $a \rightarrow b_s$ satisfies the two criteria 10 and 11.

This algorithm provides a top-down search of rules in $\mathcal{H}_2$, and then explores all branches of the hierarchy. We choose to stop the descent in a branch if $\forall b'_x \leq_2 b_x, \varphi(a \rightarrow b'_x) < \varphi_r$. For a rule $x \rightarrow y$, a property of implication intensity defines $x \cup y$ as the best specialization of the consequent. We exploit this property in order to avoid the evaluation of all rules $a \rightarrow b'_x$.

The describing search method does not consider the roots of hierarchies because all selected terms are associated to root-concepts. The implication intensity value of such rules (i.e. rules which contain root-concepts) is either undefined or equals to 0.

Global variable :
  φ_r : The Implication Intensity threshold
Inputs :
  a : a concept of $\mathcal{H}_1$.
  b : a concept of $\mathcal{H}_2$.
  φ_{max} : The best value $\varphi(a \rightarrow b_p)$ with $b \leq b_p$
Input/Output variables :
  $\mathcal{B}_{current}$: the list of "current" concepts taking from $\mathcal{H}_2$.
  $ruleList$: the list of selected rules.
return value :
  The value φ of the best rule $a \rightarrow b_x$ with $b_x \leq b$

Function specializeConsequent(a, b, $\mathcal{B}_{current}$,φ_{max}, $ruleList$)
Begin
  $bestChild := FALSE$
  $\varphi_{current} := \varphi(a, b)$
  $returnVal := \varphi_{current}$
  If $(\varphi_{current} < \varphi_r)$ **then**
    $\varphi' := \varphi(a, a \cap b)$
    If $(\varphi' < \varphi_r)$ **then**
      return $\varphi_{current}$
    EndIf
  EndIf
  ForEach $child \in children(b)$ **do**
    $\varphi_{child} := specializeConsequent(a, child, \mathcal{B}_{current}, ruleList)$
    If $(\varphi_{child} \geq \varphi_{current})$ **then**
      $bestChild := TRUE$
      $\mathcal{B}_{current} := \mathcal{B}_{current} - \{b\}$
      If $(returnVal < \varphi_{child})$ **then**
        $returnVal := \varphi child$
      EndIf
    EndIf
    If $(\varphi_{current} > \varphi_r)$ and $\rceil(bestChild)$ and $(\varphi_{current} \geq \varphi_{max})$ **then**
      $ruleList := ruleList \cup (a \rightarrow b)$
      $\mathcal{B}_{current} := \mathcal{B}_{current} \cup \{b\}$
      $\varphi_{max} := \varphi_{current}$
    EndIf
  EndDo
  return $returnVal$
End

Figure 5. Algorithm specializing consequent

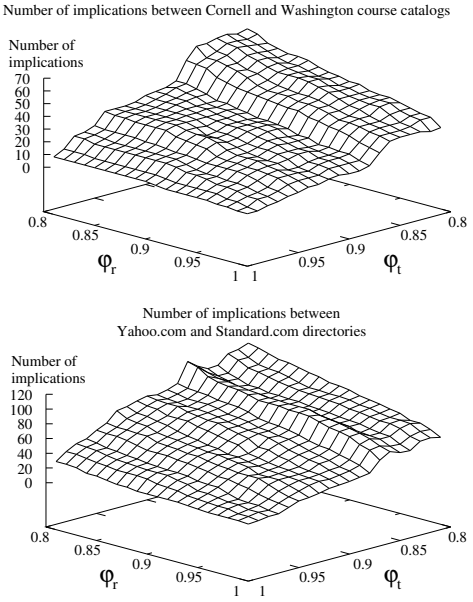


Figure 6. Influence of threshold values over the amount of selected rules

7 Experiments

In this section, we choose to compare our method results with a benchmark which is manually matched. We do not confront our results with those provided by other approaches because among the few extensional approaches proposed in the litterature, none are asymmetric. This fact implies that we must "symmetrisize" our results before making any experimental comparison with other extensional approaches. This "symmetrization" introduces a strong bias.

7.1 The analysed data

We experimented our algorithms on two benchmarks proposed in [7]. The first benchmark "Course catalog" describes courses which are proposed at the Cornell and Washington universities. The courses descriptions are hierarchically organised. These two hierarchies contain respectively 166 and 176 concepts to which are associated 4360 and 6957 textual course descriptions. The second benchmark "Company profiles" is issued from the web directories Yahoo.com and Standard.com. These hierarchies describe respectively 13634 and 9504 companies. These company profiles are respectively organised into 115 and 333 sectors and industries.

7.2 Results

We perform both a quantitative test and a qualitative test over the datasets. First, we vary the term threshold φ_t and the rule threshold φ_r in order to analyse their influence over the amount of discovered rules. Secondly, we compare our experimental results with manual matching reference. We vary the thresholds values from 0.8 to 1 for the two tests.

The two graphs of figure 6 shows that the terms selection threshold φ_t has a greater influence than the rule selection threshold φ_r. For example, a φ_t increase of 0.1 unit entails a decrease of 2.15 rules while the same increase of 0.1 unit for φ_r only entails a decrease of 1.2 rules.

Secondly, we perform the qualitative test of precision and recall with the "course catalog" dataset. We compare the results produced by our approach with a reference matching pair set provided by [7]. However their relations are symmetric while ours are asymmetric. Consequently we only retain equivalence relations from ours results (If $a \rightarrow b$ and $b \rightarrow a$ then $a \leftrightarrow b$). Finally, we built the two following graphs respectively representing the evolution of the precision value and the evolution of the recall value in function of chosen threshold values φ_t and φ_r. These two measures issued from information retrieval are defined as follows: let us considers F the set of matching pairs found using our approach and R the set of "reference" matching pairs. The precision ($precision = card(F \cap R)/card(F)$) measures the share of real matching pairs among all found ones. The recall ($recall = card(F \cap R)/card(R)$) measures the share of real correspondances that is found.

Figure 7 shows that the term selection threshold φ_t has a greater influence than the rule selection threshold φ_r. We obtain good precision values (from 0.71 to 1). Nevertheless, the recall values are quite bad: we notice an average recall value equals to 0.29. The best value is equal to 0.54. Our method seems to be too selective.

We found two arguments explaining these recall results. The former is related to the term selection phase. Indeed, a lot of leaves-concepts of the hierarchies cannot be compared to other concepts because they do not have relevant terms selected. Leaves-concepts are not associated with a lot of documents, so it is difficult to assign relevant terms to those concepts. The latter reason, is due to the kind of studied relations. We consider in our approach implicative relations from which we deduce equivalence relations. For example, if

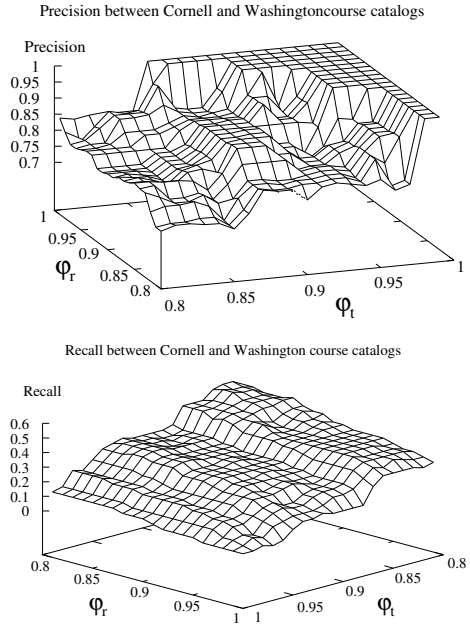


Figure 7. Thresholds values influence over the precision and the recall

we hold the term selection threshold φ_t to 0.83 and the rule selection threshold φ_r to 0.82, we obtain 24 false positives. Among them, 6 are considered as simple implications and not equivalence.

However, our method discovered some meaningful rules which are not in the manual matching reference. For example, we have obtain : From Cornell to Washington :

City and Regional Planning → Urban Planning URBDP
Cognitive Studies Program → Psychology PSYCH
Department of Aerospace Studies → Aerospace Studies A S
Electrical and Computer Engineering → Electrical Engineering E E

From Washington to Cornell :

Atmospheric Sciences ATM S → Earth and Atmospheric Sciences

Our approach is not sensitive to concept name: it led to the following rule "Cognitive Studies Program → Psychology PSYCH" that would not be found with an approach based on string similarity. The selection of relevant terms for concepts permits to take into account the semantics of the concepts.

8 Conclusion

In this paper, we proposed an extensional matching method based on the discovery of significant implication rules between concepts. Our approach takes in two hierarchies of concepts to be matched and the textual corpus indexed to these concepts. The matching task is divided into two stages: (1) the extraction and selection of relevant terms for each concepts; (2) the discovery of significant rules between concepts by using their relevant terms set. The main advantages of this method are the consideration of semantic by using binary terms contained in the corpus and the discovery of rules allowing to enhance the produced matching results only regarding similarity-based matching systems. We implemented and tested our algorithms with two real catalogs respectively related to company profiles and course catalogs. The results show that we distinguish implication and equivalence relations. We can also notice that our prototype has found several relevant relations not considered by the manual reference matching pair set.

In our point of view, and after a study of related works, our method seems really novel and unique because: (1) it works on terminological level and not on the document level, (2) it is based on asymmetric relations (which offer a stronger semantic for user), (3) it operates an original reduction of rule redundancies.

Currently, we propose an extensional individual matcher. In the near future, we will propose a schema based matcher and combine the two approaches in order to enhance the ontology matching task.

REFERENCES

[1] R. Agrawal, T. Imielinski, and A. Swami, 'Mining association rules between sets of items in large databases', in *the 1993 ACM SIGMOD international conference on Management of data*, pp. 207–216. ACM Press, (1993).

[2] J. Blanchard, P. Kuntz, F. Guillet, and R. Gras, *Implication intensity: from the basic statistical definition to the entropic version*, chapter 28, 473–485, CRC Press, 2003.

[3] S. Castano, V. De Antonellis, and S. De Capitani Di Vimercati, 'Global viewing of heterogeneous data sources', *IEEE Transactions on Knowledge and Data Engineering*, **13**(2), (2001).

[4] B. Daille, 'Conceptual structuring through term variations', in *ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, eds., F. Bond, A. Korhonen, D. MacCarthy, and A. Villaciencio, pp. 9–16, (2003).

[5] H.H. Do and E. Rahm, 'Coma - a system for flexible combination of schema matching approaches', in *the International Conference on Very Large Data Bases (VLDB '02)*, pp. 610–621, (2002).

[6] A. Doan, J. Madhavan, P. Domingos, and A. Halevy, 'Learning to map between ontologies on the semantic web', in *The Eleventh International WWW Conference*, pp. 662–673. ACM Press, (2002).

[7] A. Doan, J. Madhavan, P. Domingos, and A. Halevy, *Ontology Matching : a machine learning approach*, 397–416, Springer-Velag, 2004.

[8] F. Giunchiglia, P. Shvaiko, and M. Yatskevich, 'S-match: an algorithm and an implementation of semantic matching', in *European Semantic Web Symposium*, LNCS 3053, pp. 61–75, (2004).

[9] R. Gras et al., *L'implication statistique, une nouvelle méthode exploratoire de données*, La pensée sauvage, 1996.

[10] Y. Kalfoglou and M. Schorlemmer, 'Ontology mapping: the state of the art', *Knowedge Engineering Review*, **18**(1), 1–31, (2003).

[11] H. Liu. Montylingua: An end-to-end natural language processor with common sense, 2004.

[12] J. Madhavan, P. A. Bernstein, and E. Rahm, 'Generic schema matching with cupid', in *the International Conference on Very Large Data Bases (VLDB'01)*, pp. 49–58, (2001).

[13] S. Melnik, H. Garcia-Molina, and E. Rahm, 'Similarity flooding: A versatile graph matching algorithm and its application to schema matching', in *the 18th International Conference on Data Engineering (ICDE'02)*, pp. 117–128. IEEE Computer Society, (2002).

[14] N. Noy and M. Musen, 'Anchor-prompt: Using non-local context for semantic matching', in *the Workshop on Ontologies and Information Sharing at the International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 63–70, (2001).

[15] E. Rahm and P. A. Bernstein, 'A survey of approaches to automatic schema matching', *The VLDB Journal*, **10**(4), 334–350, (2001).

[16] P. Shvaiko and J. Euzenat, 'A survey of schema-based matching approaches', *Journal on Data Semantics IV*, **4**(LNCS 3730), 146–171, (2005).

[17] G. Stumme and A. Maedche, 'FCA-MERGE: Bottom-up merging of ontologies', in *IJCAI*, pp. 225–234, (2001).

Annexe F

Curriculum Vitae

GUILLET Fabrice

Ecole polytechnique de l'université de Nantes
rue Christian Pauc - La Chantrerie - BP 60601
44306 Nantes Cedex 3.
Tél : 02 40 68 30 92, Fax: 02 40 68 32 32
Courriel : Fabrice.Guillet@polytech.univ-nantes.fr

Né le 30 Mai 1965 à Nantes, 41 ans;
de Nationalité Française,
marié, 1 enfant,
libéré des obligations militaires.

Maître de Conférences en Informatique

Docteur en informatique - Ingénieur en Télécommunications

Table des Matières

Partie I : Synthèse

1. PARCOURS

Coordonnées

Nom, Prénom : Guillet Fabrice

Date de naissance : 30 mai 1965, à Nantes (44)

Adresse personnelle : 43 rue du Clos Siban – 44800 Saint-Herblain

Téléphone personnel : 02 40 28 90 61

Adresse professionnelle : Polytech'Nantes -Ecole polytechnique de l'université de Nantes - rue Christian Pauc - La Chantrerie - BP 60601 - 44306 Nantes Cedex 3

Téléphone professionnel : 02 40 28 90 61, **Fax:** 02 40 68 32 32,

Courriel : Fabrice.Guillet@univ-nantes.fr

Emplois

Depuis 1997 : *Maître de Conférences*, Département Informatique, Ecole Polytechnique de l'université de Nantes (Polytech'Nantes, ex-IRESTE). Recherches en Extraction et Gestion de Connaissances, équipe COnnaissances et Décision (COD), Laboratoire d'Informatique de Nantes Atlantique (LINA CNRS FRE 2729, ex-IRIN).

1995-1997 : *ATER*, IRESTE, Université de Nantes. Recherches en ECD, équipe Connaissances Informations Données (CID), Institut de recherche en Informatique de Nantes (IRIN).

1992-1995 : *Vacataire*, Département Intelligence Artificielle et Sciences Cognitives (IASC), Ecole Nationale Supérieure des Télécommunications de Bretagne (ENST Bretagne). Recherches en acquisition de connaissances et aide à la décision.

1989-1991 : *Enseignant Volontaire Service National Actif* (*VSNA)*, traitement du signal et télécommunications, Ecole Royale de l'Air de Marrakech, en collaboration avec l'Ecole de l'Air de Salon de Provence.

Diplômes

1995 : *Doctorat* en informatique, université de Rennes I. Convention CIFRE entre THOMSON-CSF et l'ENST Bretagne. Directeur : J.-P. Barthélemy.

1989 : *Ingénieur*, Ecole Nationale Supérieure des Télécommunications de Bretagne.

1987 : *Maîtrise* Sciences et Techniques en traitement de l'information, université de Rennes I.

1985 : *DUT* génie électrique et informatique industrielle, université de Rennes I.

2. RESPONSABILITES

Cette partie résume la partie saillante des responsabilités assurées en recherche, pédagogie et contrats. Celles-ci sont détaillées dans les sections *Responsabilités en recherche* (Partie III :3, page 39), *Responsabilités scientifiques* (Partie III :8.1, page 56), *Responsabilités pédagogiques* (Partie IV :1, page 65).

Responsabilités en recherche – niveau équipe

- [02-04] *Co-responsable* du thème "Extraction et Gestion des Connaissances" de l'équipe "Connaissances Informations Données" (CID) de "l' Institut de Recherche en Informatique de Nantes" (IRIN, UPRES 2729), (8 permanents, 4 associés, 7 doctorats). Ce thème est devenu une équipe en 2004.
- [99-…] *Co-responsable* de 2 axes de l'équipe COnnaissances et Décision (COD), Laboratoire d'Informatique de Nantes Atlantique (LINA, CNRS FRE 2729) :
  - Mesure de la Qualité des Connaissances (4 permanents, 3 doctorants.
  - Gestion et Ingénierie des Connaissances (4 permanents, 1 doctorant).
- [01-…] *Membre élu* de la *commission de spécialistes*, section 27. [01], puis réélu [04]

Responsabilités en recherche – niveau national

- [03-…] *Chargé de communication* de *l'association "Extraction et Gestion des Connaissances"* (EGC). (*Membre élu*, 1er mandat [02-05], puis 2ème mandat [depuis 05]). Fondateur de l'association. Fonctions : coordination inter-conférences, gestion des parrainages, du site web, des adhérents, et de la liste de diffusion (1000 abonnés).
- [03-…] *Membre permanent du comité de pilotage* du cycle de conférences francophones "Extraction et Gestion des Connaissances" (EGC).
- *Responsabilités éditoriales* – maître d'oeuvre de l'édition de deux numéros spéciaux de revues nationales comportant un comité de rédaction :
  - (1) H. Briand et F. Guillet (eds), revue Extraction des Connaissances et Apprentissage (ECA), vol. 1, n° 1-2, 2001. Hermès Science Publication.
  - (2) H. Briand, M. Sebag, R. Gras, F. Guillet (eds), revue Nationale des Technologies de l'Information (RNTI), E1, 2004, Cépaduès.
- [01] *Président* du *comité d'organisation* de la 1ère conférence francophone Extraction et Gestion des Connaissances, EGC, Nantes 2001.
- [05-06] *Président du Comité de Programme* de 2 ateliers «Qualité des Données et des Connaissances », QDC/EGC, Paris 2005, Lille 2006.
- [02-03] *Co-animateur* du *groupe de travail GafoQualité* (30 chercheurs, 10 équipes) de l'Action Spécifique AS GafoDonnées - STIC - CNRS. Animation du groupe, organisation de 4 séminaires, et co- rédaction d'une synthèse nationale (1 rapport CNRS)

Responsabilités en recherche – niveau international

- [02-…] *Participation à la création et chargé de la coordination internationale* pour le Master "Extraction de Connaissances dans les Données" (Master ECD), coordination avec la Roumanie (Politehnica Bucarest) et le Vietnam (University of Cantho), selon 2 activités principales :
  - (1) responsable administratif et pédagogique des échanges internationaux,
  - (2) responsable du système de téléenseignement par visioconférence réparti sur 4 sites.
- [06-…] *Participation à la création et chargé de la coordination internationale* pour le projet de Master MDM&KD, extension européenne du Master ECD. Participation à la création de ce master et extension de l'activité précédente à l'université du Piémont Oriental, Italie.
- [00-…] *Invitation de chercheurs internationaux* : invitation et accueil de 7 chercheurs internationaux, dont G. Piatetsky-Shapiro en 2005.
- *Responsabilités éditoriales* – maître d'oeuvre de l'édition de deux livres de chapitres comportant un comité de rédaction international :
  - [06] F. Guillet and H. Hamilton (eds), « Quality measures in Data Mining », 2007, Springer. Parrainé par G. Piatetsky-Shapiro, à paraître.
  - [04] R. Gras, E. Suzuki, F. Guillet and F. Spagnolo (eds), « Statistical Implicative Analysis », en cours de construction, édition prévue pour 2007.
- [2005] *Session invitée*: F. Guillet and P. Lenca « Knowledge Quality », ASMDA conference, Brest 2005.

Responsabilités Pédagogiques

- [01-…] *Responsable des Relations Internationales* du Département Informatique, Polytech'Nantes. J'assure intégralement la gestion administrative, logistique, financière et pédagogique de trois tâches principales :
  - (1) les stages ingénieurs (35-50/an) et Master (4-6/an) à l'étranger,
  - (2) les mobilités étudiants (10-15 semestres/an) et enseignants (1-2 x 1 mois invité/ans), et les invitations (2-4 visiteurs/an),
  - (3) les conventions inter-universitaires avec les universités étrangères (40 au total, dont 3 depuis 2004, et 2 en cours).
- [00-…] *Création de formations* – implication forte dans la création, l'enseignement et la responsabilité pédagogique de modules :
  - (1) du cursus ingénieur « Informatique Décisionnelle », option lourde, 3ans, 2000h, du département Informatique de polytech'Nantes [02-…],
  - (2) du Master de recherche « Extraction des connaissances dans les Données » [00-…]
  - (3) du Master européen « Master in Data mining and Knowledge Discovery » [06-…].
- [05-…] *Equipements pédagogiques* : Responsable de 3 équipements pédagogiques lourds (initiateur et porteur du projet, recherche du financement, déploiement et gestion de l'équipement) :
  - (1) salle « réalité virtuelle » (financement régional 15 k€),
  - (2) salle « visioconférence » (financement interne 10 k€).
  - (2) salle « Informatique Décisionnelle » (financement régional 20 k€).
- [97-…] *Conception de modules* – conception des programmes pédagogique, gestion des intervenants, et participation aux enseignements : **3** (Master), **6** (Ingénieur), **1** (IUP [99]), **1** (Formation continue [97-02]).
- [01-…] Membre de la *commission matériels et logiciels* du département Informatique.
- [04-…] Membre nommé au *conseil de département*

Responsabilités sur Projets et Contrats – Niveau national

- [04-06] *Co-responsable* pour l'équipe COD du projet "Groupe d'Agents Collaboratifs Emotionnels" (RIAM GRACE). Maître d'oeuvre du projet pour COD, gestion d'une équipe projet (10 stages, 4 CDD, 40 h-mois), rédaction du rapport d'activité COD.
- [98-04] *Co-responsable* de 2 *contrats lourds* avec production d'un logiciel pour des entreprises :
  - [98-01] Samanta avec LA POSTE, Concepteur et maître d'oeuvre du projet, gestion d'une équipe projet (16 stages, 70 hmois), campagne de recueil de connaissances.
  - [01-04].Athanor avec PerformanSE SA [00-04]. Maître d'oeuvre du projet, gestion d'une équipe projet (16 stages, 60 hmois), avec transfert technologique.
- [02-04] *Co-responsable* pour l'équipe COD de l'axe MEC du projet quadriennal état-région "Connaissances Objets Modélisation" (CPER COM). Animation scientifique et rédaction des rapports d'activité

Responsabilités sur Projets et Contrats – Niveau International

- *Membre* du *réseau d'excellence européen* : Interoperability Research for Networked Enterprises Applications and Software (Interop NoE), participation à 3 workpackages [03-06].

Autres

- *Membre de 5 groupes de recherche* : AS GafoDonnées [02-03], AS Topik [03-04], HCP Euro Group [03-…], GDR I3 [02-…], GDR Fouille de données [04-…].
- Membre du *conseil d'administration* de l'association *Polyteco*. [01-…]
- Membre de diverses *commission ponctuelles* de réflexions : semestrialisation [00], contrats industriels [01], harmonisation des stages [03], …
- Responsable de l'administration du serveur recherche de l'équipe CID [00-02].

3. RECHERCHE

Publications

Catégorie	*Avant 2002*	*2002-2005*	*Après 2005*	***Total***
Livres et Chapitres Internationaux		1	2	**3**
Livres et Chapitres Nationaux	1	1		**2**
Revues Internationales			1	**1**
Revues Nationales	4	13	4	**21**
Conférences Internationales	8	10	8	**26**
Workshop Internationaux	3	1	1	**5**
Conférences invitées		2		**2**
Conférences Nationales	5	2		**7**
Conférences Nationales à publication restreinte	3	7	5	**15**
Autres (colloques, séminaires, rapports)	8	12	3	**23**
				105

Encadrements

- Thèses : **7** (3 soutenues, 4 en cours)
- DEA/Masters : **18**
- Jurys de thèses externes : **2**

Recherche Contractuelle

- Contrats de recherche sur financements publics : **5** (NoE, CPER, RIAM, 2 Fondation R&E Fondation de France)
- Contrats de recherche sur financements privés : **2** (La Poste, PerformanSE SA)
- Bourses de thèse : **7** (2 CIFRE, 1 Bourse Gouv. Vietnamien, 2 Fondation R&E, 1 sur contrat avec La Poste, 1 bourse régionale).

Logiciels et transfert technologique

- **2** *logiciels lourds* ayant nécessité plus de 3 ans de développement et ayant abouti à un transfert technologique (SAMANTA avec La Poste, et ATHANOR avec Performanse SA).
- **3** *prototypes* issus de nos travaux de recherche et téléchargeables sur la toile (ARVAL, ARVIS, ARQAT).

Rayonnement national

- *Membre fondateur* de l'association Extraction et Gestion des Connaissances [02-…]
- *Comité de pilotage* conférence Extraction et Gestion des Connaissances (EGC) [03-…]
- *Animation* de *groupes de travail* : **1** groupe GafoQualité AS GafoDonnées [02-03]
- *Comités de lecture de revues nationales* : **4**, dont 3 num. spé., et 1 comme relecteur.
- *Comités de Programme nationaux* : **20**, dont **2** comme *président*, et **1** comme relecteur
- *Comités d'Organisation nationaux* : **6**, dont **3** comme *président*
- *Conférences invitées nationales* : **1** tutoriel
- Participation à des *groupes de travail nationaux* : **4**
- Présidence de *session* dans des conférences nationales : **5**.

Rayonnement international

- *Comités de lecture de revues internationales* : **3** comme relecteur
- *Comités de Programme internationaux* : **19**, dont **ICDM** (conf. Int. Phare en data mining), et dont **7** comme relecteur
- *Comités d'Organisation internationaux* : **1** workshop
- *Conférences invitées internationales* : **1** session invitée
- Participation à des *groupes de travail Européens* : **2**, dont **1** Réseau d'Excellence *Interop NoE* [03-06]
- Présidence de *session* dans des conférences : **2**.

Coopérations Internationales

- **7** chercheurs internationaux invités, dont G. Piatetsky-Shapiro en 2005.
- **8** collaborations principales en recherche : (1) Politehnica Univ. Bucarest Romania, (2) Cantho Univ. Vietnam, (3) Univ. Piémont Italy, (4) Univ. Regina Canada, (5) Univ. Laval Canada, (6) Univ. Yokohama et Kyushu Japan, (7) UTS Sidney Australia, (8) Univ. Palermo Italy;
- **12** collaborations ponctuelles en recherche ;
- collaboration avec plus de **30** universités supplémentaires pour les échanges contractualisés d'étudiants (ERASMUS, AUF, EGIDE).

4. ENSEIGNEMENT

Encadrements de projets

- Mémoires Cnam : **5** (Cycle C, ingénieur en informatique, 9-12 mois) [97-…]
- Probatoires Cnam : **4** (Cycle C, ingénieur en informatique, 5 semaines) [00-…]
- Stages Master informatique : **17** (Master SAD et Master ECD, 5 mois) [97-…]
- Stages ingénieur, Polytech'Nantes : **5** (3eme année, bac+5, 5 mois) [97-…]
- Projets 2eme année, Polytech'Nantes : **3 à 4/an** (8 mois, 300 h) [97-…]
- Projets 3eme année, Polytech'Nantes : **2 à 3/an** (4mois, 150 h) [97-…]
- Stages 3eme année (tuteur) : **3 à 4/an** (5mois) [97-…]
- Projets formation continue DUTIL, Polytech'Nantes : **1 à 2/an** (6 mois, 200 h) [95-02]
- Autres projets ingénieur : **3** ENST Bretagne [92-95] et **4** Ecole Royale de l'Air (ERA, 2 mois, 150 h) [89-91]

Enseignements dispensés

Module	Niveau	Lieu	Type	Années	H/an
Théorie du signal	Ingénieur, Bac+4	ERA	Cours	89-91	50
Théorie du signal	Ingénieur, Bac+3	ERA	Cours/TP	89-91	50/50
Smalltalk	Ingénieur, Bac+4	ENST Br	TP	92-95	20
Smalltalk	Mastère, Bac+6	ENST Br	Cours/TP	92-95	20/20
Langages et compilateurs	Ingénieur, Bac+4	ENST Br	TP	92-95	20
Smalltalk	Ingénieur, Bac+3	Polytech'Nantes	TP	95-97	12
Unix et Syst. Exploitation	Ingénieur, Bac+3	Polytech'Nantes	TP	95-97	50
Langage C	Stage FC, Bac+4	Polytech'Nantes	Cours/TP	95-97	20/15
Unix	Stage FC, Bac+4	Polytech'Nantes	Cours/TP	95-97	20/10
Unix	Dutil, Bac+4	Polytech'Nantes	Cours/TP	95-97	10/14
Algorithmique et langage C	Dutil, Bac+4	Polytech'Nantes	Cours/TP	95-97	10/14
Smalltalk	Dutil, Bac+4	Polytech'Nantes	Cours/TP	95-02	10/10
Algorithmique et langage C	Dutil, Bac+4	Polytech'Nantes	Cours/TP	95-02	10/20
Java	Dutil, Bac+4	Polytech'Nantes	Cours/TP	97-02	6/9
Architecture des Systèmes	Ingénieur, Bac+3	Polytech'Nantes	Cours/TP	97-…	40/50
Programmation objet : java, C++	Ingénieur, Bac+3	Polytech'Nantes	TP	97-…	40
Extraction de Connaissances (ECD)	IUP GIS, Bac+4	IUP Vannes	Cours	99	12
IHM	Ingénieur, Bac+3	Polytech'Nantes	Cours/TP	97-01	8/12
Traitement de l'information	Ingénieur, Bac+3	Polytech'Nantes	TD	97-00	12
Mesure de qualité en ECD	Ingénieur, Bac+5	ENST Bretagne	cours	00-02	3
Réseaux et Télécommunications	Ingénieur, Bac+3	Polytech'Nantes	TD	00-02	12
Fouille de données - ECD	Ingénieur, Bac+5	Polytech'Nantes	Cours/TP	01-05	18/6
Parallélisme - Clusters	Ingénieur, Bac+5	Polytech'Nantes	Cours/TP	01	9/9
Réseaux - Sécurité	Ingénieur, Bac+5	Polytech'Nantes	Cours	01	9/9
Extraction de Connaissances (ECD)	Ingénieur, Bac+4	Polytech'Nantes	Cours/TP	04-…	12/6
Gestion des Connaissances	Ingénieur, Bac+3	Polytech'Nantes	Cours/TP	03-…	12/6
Web Sémantique	Ingénieur, Bac+5	Polytech'Nantes	Cours/TP	05-…	3/6
Qualité des connaissances	DEA ECD	Polytech'Nantes	Cours	00-02	5
Extraction et Gestion des Connaissances	DEA ECD	Polytech'Nantes	Cours	02-04	6
Gestion et déploiement des Connaissances	Master ECD	Polytech'Nantes	Cours	04-…	12
Extraction de Connaissances (ECD)	DEA SAD	Univ. Nantes	Cours	02-04	6
ECD – fouille de données, qualité	Master SAD	Univ. Nantes	Cours	04-…	9

Partie II : Curriculum Vitae

1. ETAT-CIVIL

Nom, Prénom : Guillet Fabrice

Date de naissance : 30 mai 1965, à Nantes (44)

Nationalité : Française

Etat-civil : Marié, 1 enfant

Adresse personnelle : 43 rue du Clos Siban – 44800 Saint-Herblain

Tél. personnel : 02 40 28 90 61

Adresse professionnelle :

Ecole polytechnique de l'université de Nantes - Polytech'Nantes

La Chantrerie

Rue Christian Pauc

BP 60601

44306 Nantes Cedex 3

Tél. professionnel : 02 40 28 90 61

Fax : 02 40 68 32 32

Courriel : Fabrice.Guillet@univ-nantes.fr

2. EMPLOIS

Depuis 1997 : *Maître de Conférences* - Section CNU 27 Informatique - Département Informatique, Ecole Polytechnique de l'université de Nantes.

1995 - 1997 : *ATER*, IRESTE, Université de Nantes.

1992 - 1995 : *vacataire*, Département Intelligence Artificielle et Sciences Cognitives, Ecole Nationale Supérieure des Télécommunications de Bretagne.

1989 - 1991 : *enseignant VSNA*, traitement du signal et télécommunications, Ecole Royale de l'Air de Marrakech, en collaboration avec l'Ecole de l'Air de Salon de Provence.

3. DIPLOMES

1995 - Doctorat : Doctorat nouveau régime, spécialité informatique, université de Rennes I. Convention CIFRE entre THOMSON-CSF et l'ENST de Bretagne. Directeur : J.-P. Barthélemy.

Titre : Contributions à la maîtrise de qualité d'un processus industriel par apprentissage des stratégies de contrôle de l'opérateur.

Mots clés : sciences cognitives, processus heuristique de décision, acquisition de connaissances, contrôle de processus, maîtrise de la qualité.

Jury :

- Alain Dussauchoy (rapporteur), professeur, Université de Lyon I
- Etienne Mullet (rapporteur), professeur, EPHE, Paris
- Jean-Pierre Barthélemy (directeur), professeur, ENST Bretagne
- Jean-Pierre Banâtre (examinateur), professeur, Université de Rennes I
- Bruno Leclerc (examinateur), HdR, EHESS, Paris
- Fabrice Lainé (co-encadrant), Thomson-CSF
- Jean-Luc Voirin (invité), Thomson-CSF

1989 - Ingénieur : Ingénieur de l'Ecole Nationale Supérieure des Télécommunications de Bretagne.

1987 - Maîtrise : Maîtrise Sciences et Techniques en traitement de l'information, université de Rennes I.

1985 - DUT : DUT génie électrique et informatique industrielle, université de Rennes I.

Partie III : Recherche

1. SYNTHESE DES TRAVAUX DE RECHERCHE

1.1. Introduction

En synthèse, l'ensemble de mes travaux caractérisés par le mot clé « connaissances » s'articule autour d'un axe sous-jacent « aide à la décision ». Cette axe a été décliné dans deux domaines de recherches : l'extraction de connaissances dans les données, et la gestion et l'ingénierie des connaissances.

1.2. Extraction de Connaissances dans les Données et règles d'association

L'extraction de connaissances dans les données (ECD) est un domaine de recherche multidisciplinaire, né dans les années 80 avec l'émergence des bases de données volumineuses. Il a été identifié par le MIT comme l'une des 10 technologies émergentes du 21ème siècle1, et reconnu par l'ACM à travers la création d'un groupe de recherche international (SIG–KDD) animé par G. Piatetsky-Shapiro. L'ECD a pour objectif l'extraction de connaissances auparavant inconnues et potentiellement utiles au sein de grands volumes de données : « Knowledge Discovery is the non-trivial extraction of implicite, previously unknown, and potentially useful information from data2 ». Ce domaine de recherche par essence multidisciplinaire s'est appuyé à l'origine sur les bases de données, puis a rapidement nécessité une étroite coopération avec l'apprentissage automatique, les statistiques et l'analyse de données, la visualisation, et l'aide à la décision.

L'ECD propose une méthodologie fondée sur un processus de transformation des données vers les connaissances, qui se décompose en trois étapes majeures :

1. La localisation, la sélection et le prétraitement des gisements de données ;
2. La découverte des connaissances à l'aide modèles prédictifs ou descriptifs, supervisés ou non supervisés ;
3. Le post-traitement des connaissances découvertes et leur validation par un décideur/expert des données.

Mes/nos travaux de recherche se situent sur les phases aval (2 et 3) du processus d'ECD, au plus proche du décideur. Plus particulièrement, je me suis intéressé au modèle des règles d'association (RA) qui permet de découvrir sans connaissances préalables des tendances implicatives au sein des données. Cet avantage d'une découverte non supervisée permise par les RA, est contrebalancé par un inconvénient majeur, celui de délivrer une quantité prohibitive de règles, qui nécessite une phase de post-traitement adaptée afin de devenir intelligible à un décideur. De mon/notre point de vue une solution réside dans l'hybridation de trois approches envisagés séparément dans la littérature :

1. Mesures de qualité : pour sélectionner les règles potentiellement intéressantes à l'aide de mesures adaptées (interestingness measures), et par l'élimination des redondances ;
2. Représentations graphiques : pour permettre la visualisation des règles potentiellement intéressantes ;
3. Interactivité : pour permettre à l'utilisateur d'exprimer des contraintes sur les règles afin de

[1] *MIT Technology Review*, 2001.
[2] W.J. Frawley, G. Piatetsky-Shapiro, et C.J. Matheus. Knowledge discovery in databases: an overview. *AI Magazine*, Fall: 57-70, 1992.

cibler celles qui l'intéressent.

Nous avons ainsi proposé des méthodes d'élimination des règles redondantes et réduction des variables. Puis, nous avons étudié et comparé les mesures de qualité de règles, et avons développé des mesures originales. Enfin, considérant le problème selon une perspective d'aide à la décision dans la quelle les représentations graphiques intelligibles et l'interactivité avec l'utilisateur jouent un rôle majeur, nous avons proposé deux solutions originales et *anthropocentrées* combinant les 3 approches précédentes et intégrant très fortement le décideur dans le processus de découverte.

1.3. Mesures de Qualité : quantifier l'intérêt d'une règle d'association

Partant d'une structure de données croisant individus et variables binaires (transactions), généralement issue d'un SGBD relationnel, les règles d'associations permettent de découvrir des tendances implicatives entre combinaisons de variables (itemsets). Les premiers travaux sur les règles d'association[3] ont été restreints à l'utilisation de mesures d'intérêt aux vertus essentiellement algorithmiques : le support et la confiance. Afin d'améliorer la sélection des meilleures règles, de nombreuses mesures complémentaires ont ensuite été proposées dans la littérature[4]. [5]Freitas[6] distingue les mesures subjectives qui intègrent les buts/connaissances du décideur, et les mesures objectives qui sont des indicateurs statistiques évaluant la contingence d'une règle.

Dans le contexte des mesures objectives, et en nous appuyant sur les travaux précurseurs sur l'analyse statistique implicative de R. Gras[7], et H. Briand et L. Fleury[8], nos apports ont été les suivants :

1. Nous avons proposé plusieurs extensions de la mesure statistique d'intensité d'implication selon les contextes : de données volumineuses[9] et de données de séquences[10], de variables ordinales[11] et floues, de réduction de variables[12] et d'élimination des redondances[13], et enfin de cohésion de classes de règles[14],.
2. Nous avons développé trois mesures de qualité originales : deux mesures entropiques EII[9]

---

[3] R. Agrawal, T. Imielinsky, et A. Swami. Mining association rules between sets of items in large databases. In Proc. of the *ACM SIGMOD'93*, pages 207-216, 1993.
[4] Tan, P.-N., V. Kumar, et J. Srivastava. Selecting the right objective measure for association analysis. *Information Systems 29(4)*, 293–313, 2004.
[5] Tan, P.-N., V. Kumar, et J. Srivastava. Selecting the right objective measure for association analysis. *Information Systems 29(4)*, 293–313, 2004.
[6] A. A. Freitas. On rule interestingness measures. *Knowledge-Based Systems 12(5-6)*, 309–315, 1999.
[7] R. Gras. *L'implication statistique - Nouvelle méthode exploratoire de données*. La Pensée Sauvage éditions, 1996.
[8] Fleury L., « Découverte de connaissances pour la gestion des ressources humaines », Thèse de doctorat, Université de Nantes, 1996
[9] J. Blanchard, P. Kuntz, F. Guillet et R. Gras. Implication intensity: From the basic definition to the entropic version, Chap. 28. In *Statistical Data Mining and Knowledge Discovery*, pages 475-493, CRC Press-Chapman & al., 2003.
[10] J. Blanchard, F. Guillet, H. Briand. L'intensité d'implication entropique pour la recherche de règles de prédiction intéressantes dans les séquences de pannes d'ascenseurs. *Extraction des Connaissances et Apprentissage (ECA)*, vol. 1, n° 4, 77-88, 2002. Hermès.
[11] S. Guillaume et F. Guillet. Une généralisation des règles d'association par l'intensité d'implication ordinale. Actes des *7èmes Journées de la Société Francophone de Classification*, pages 77-86, Nancy, France, Septembre 1999.
[12] R. Couturier, R. Gras, F. Guillet. Reducing the number of variables using implicative analysis. In Proc. of the *Int. Federation of Classification Societies, IFCS'2004*, pp 277-285, Chicago, USA, July 15-18, 2004. Springer Verlag.
[13] R. Lehn, F. Guillet, H. Briand. Qualité d'un ensemble de règles : élimination des règles redondantes. *Revue Nationale des Technologies de l'Information (RNTI)*, E1, pp 141-168, 2004, Cépaduès.
[14] R. Gras, J. David, J.-C. Régnier, F. Guillet. Typicalité et contribution des sujets et des variables supplémentaires en Analyse Statistique Implicative. *Revue Nat. des Techn. de l'Information (RNTI)*, 2006, E6, 359-370, 2006. Cépaduès.

TIC[15] et une mesure statistique IPEE[16].

3. Enfin, nous avons proposé un ensemble conséquent de travaux de synthèse sur les mesures de qualité. D'une part, nous recensé et proposé un classement original[17] de 40 mesures d'intérêt, sur lesquelles nous avons mené une série d'études comparatives à l'aide de graphes corrélatifs[18]. Ces travaux ont été implémentés dans la plateforme ARQAT[19]. D'autre part, nous avons produit deux ouvrages de synthèse sur les mesures de qualité en ECD : une synthèse de travaux à l'échelle nationale[20] issu de l'AS GafoDonnées STIC-CNRS, et une synthèse internationale[21] dans un ouvrage collectif publié par Springer.

1.4. Processus anthropocentré de fouille de règles : utilisateur, visualisation et interaction

Partant du constat qu'en dépit de la réduction réalisée par les mesures de qualité, le nombre de règles produites demeure prohibitif, nous avons choisi d'intégrer plus fortement le décideur dans le processus d'ECD, lui restituant ainsi son rôle d'acteur majeur intégré au sein même du processus. Ce choix a été doublement motivé, considérant d'une part le nécessaire recentrage sur l'utilisateur prôné par Brachman et Anand[22], et d'autre part les travaux que j'avais menés lors de ma thèse à l'ENST de Bretagne sur les systèmes anthropocentrés d'aide à la décision au sein du Projet JADAR[23] sous la direction de Jean-Pierre Barthélemy. Ce glissement du post-traitement vers une approche interactive d'aide à la décision, où l'utilisateur joue un rôle d'« heuristique » dirigeant la fouille, nous a amené à intégrer trois composantes originales dans notre approche de « processus anthropocentré de fouille de règles » :

1. Une représentation graphique adaptée aux règles d'association, compatible avec les contraintes cognitives du décideur ;
2. Des opérateurs d'interaction avec la représentation visuelle supportant les besoins de découverte du décideur ;
3. Des algorithmes de fouilles ad hoc, connectés à la base de données, et pilotés par le décideur à travers les opérateurs d'interaction.

Le principe sous-jacent à cette approche, et issu des contraintes cognitives du décideur, a été dénommé *ciblage de règles* (rule focusing). Il consiste à passer d'une recherche globale (sur la totalité des règles), à une suite de recherches locales (sur un sous-ensemble de règles proches), la progression de t à t+1 dans la suite étant déterminée par des opérateurs

---

[15] J. Blanchard, F. Guillet, R. Gras, H. Briand. Using Information-theoretic Measures to Assess Association Rule Interestingness. In Proc. of the *5th IEEE Int. Conf. on Data Mining, ICDM'2005*, pp. 66-73, 2005, IEEE Press.

[16] J. Blanchard, F. Guillet, H. Briand, R. Gras. Une version discriminante de l'indice probabiliste d'écart à l'équilibre pour mesurer la qualité des règles. *Revue Quaderni di Ricerca in Didattica*, n°15, pp 131-138, 2005.

[17] J. Blanchard, F. Guillet, H. Briand, R. Gras. Une version discriminante de l'indice probabiliste d'écart à l'équilibre pour mesurer la qualité des règles. *Revue Quaderni di Ricerca in Didattica*, n°15, pp 131-138, 2005.

[18] X.-H. Huynh, F. Guillet, J. Blanchard, P. Kuntz, R. Gras and H. Briand. A graph-based clustering approach to evaluate interestingness measures : a tool and a comparative study. *Quality Measures in Data Mining, Studies in Computational Intelligence*. 2006. Springer. To appear

[19] X.-H. Huynh, F. Guillet, H. Briand. ARQAT: An Exploratory Analysis Tool For Interestingness Measures. In Proc. of the *11th Int. Symposium on Applied Stochastic Models and Data Analysis, ASMDA'2005*, pp 334-344, 2005.

[20] H. Briand, M. Sebag, R. Gras, F. Guillet (eds). Mesures de qualité pour la fouille de données. Numéro spécial, *Revue Nationale des Technologies de l'Information (RNTI)*, E1, 2004, Cépaduès.

[21] F. Guillet and H. Hamilton (eds). Quality Measures in Data Mining. *Studies in Computational Intelligence*. 2006. Springer. To appear.

[22] R.J. Brachman et T. Anand. The process of knowledge discovery in databases: a human-centered approach. In *Advances in Knowledge Discovery and Databases*, pages 37-58, 1996.

[23] Jugement, Aide à la Décision et Apprentissage de Règles

d'interaction déclenchés sur la représentation visuelle à l'instant t selon les desiderata de l'utilisateur.

Cette approche fortement dynamique, engendre de multiples propriétés sympathiques. La représentation visuelle n'a plus besoin de supporter l'intégralité des règles, ce qui améliore l'intelligibilité des résultats présentés. Les opérateurs incorporent une sémantique locale plus accessible à l'utilisateur et cohérente avec son activité (ex : règles plus générales/spécifiques, règles d'exception,…). Enfin, et cet apport me semble majeur, les algorithmes de fouille ad hoc deviennent locaux, ce qui a le grand avantage de rendre possible la découverte de règles d'association sur des corpus très volumineux (beaucoup de variables), là où les algorithmes globaux comme Apriori, ou les algorithmes sous contraintes, s'effondrent…

Notre approche a donné lieu à deux implémentations complémentaires.

1. Dans la première implémentation, Felix[24], l'ensemble des règles est considéré comme un immense réseau dont on ne représente qu'une partie à l'aide d'un graphe. La principale difficulté rencontrée a été de préserver la carte mentale de l'utilisateur, risquant d'être perturbée par le caractère fortement dynamique de la représentation, puisque celle-ci doit évoluer partiellement à chaque interaction.
2. Dans la deuxième implémentation, ARVIS[25],[26] , nous avons choisi de représenter chaque sous-ensemble de règle à l'aide d'une métaphore 3D orientée qualité. Chaque règle est associée à une forme graphique (une sphère posée sur un cylindre) dans un environnement tridimensionnel, et le placement ainsi que les caractéristiques graphiques de la forme/règle (dimensions, couleur) dépendent de cinq mesures de qualité caractérisant de cette règle. L'usage de la 3D permet ici d'augmenter la quantité d'information présentée sur la représentation tout en en conservant l'intelligibilité. Une extension réalisée en réalité virtuelle offre d'intéressantes perspectives d'immersion de l'utilisateur dans la représentation. Utilisée par un expert psychologue pour la découverte de règles d'association entre profils psychologiques en gestion de ressources humaines, cette seconde version a reçu un écho très prometteur.

1.5. Gestion et Ingénierie des connaissances : serveur de connaissances et ontologies

Parallèlement, stimulés par des projets émanant d'entreprises, et dans une perspective initiale plus applicative, nous nous sommes intéressés à la gestion[27] et l'ingénierie des connaissances[28] et son prolongement récent pour le web sémantique. Le lien avec l'ECD apparaît naturellement lors de la phase de déploiement des connaissances extraites qui est du ressort de la gestion et de l'ingénierie des connaissances. La gestion des connaissances a pour objet le traitement de la connaissance au sens large. Elle concerne l'acquisition, la formalisation, le stockage, la diffusion et la manipulation des connaissances et des savoir-faire généralement détenus par les acteurs (souvent experts) d'une organisation dans un domaine donné. Les deux principales limites rencontrées par les systèmes issues d'une gestion de

[24] R. Lehn, F. Guillet, P. Kuntz, H. Briand and J. Philippé. Felix : An interactive rule mining interface in a kdd process. In *Proc. of the 10th Mini-Euro Conference, Human Centered Processes, HCP'99*, pages 169-174, 1999.
[25] J. Blanchard, F. Guillet, H. Briand. A User-driven and Quality oriented Visualization for Mining Association Rules. In Proc. of the *3rd IEEE International Conference on Data Mining, ICDM'2003*, pp 493-497, 2003, IEEE Press.
[26] J. Blanchard, F. Guillet and H. Briand. Interactive Visual Exploration of Association Rules with the Rule Focusing Methodology. *Knowledge and Information Systems (KAIS)*, 2006. Spriger. ISSN: 0219-1377. To appear.
[27] J-L Ermine. *Les systèmes de connaissances*. Hermès, Paris, 1996, deuxième édition 2000.
[28] R. Dieng, O. Corby, A. Giboin, and M. Ribiere. Methods and Tools for Corporate Knowledge Management. *International Journal of Human-Computer Studies, special issue on Knowledge Management*, 51:567--598, 1999.

connaissances concernent leur déploiement dans un système d'information, souvent difficile ; et leurs capacités à évoluer afin de maintenir des connaissances à jour, souvent faibles.

Ma première activité dans le domaine a consisté à concevoir une approche « serveur de connaissances », selon une analogie avec un serveur web transposant ses services à la connaissance et bénéficiant des technologies du web, simplifiant son déploiement. J'ai implémenté cette approche dans l'outil SAMANTA[29], destiné à la gestion des diagnostics de maintenance, dans le cadre d'un contrat de 5 années avec La Poste. Ce logiciel[30] est fondé sur 3 points de vue ontologiques complémentaires : une ontologie pour les tâches de diagnostics, une ontologie décrivant la composition d'une machine de tri, une ontologie des symptômes. L'outil incorpore un éditeur d'ontologies entièrement graphique garantissant l'évolutivité du système, permet l'association à des ressources multimédia complémentaires (dont des modèles de machine de tri en réalité virtuelle), et enfin un raisonneur prolog permettant d'offrir notamment une fonctionnalité d'aide à la décision pour le diagnostic.

J'ai ensuite conçu Athanor[31], une généralisation de cette approche serveur de connaissances, incorporant une quatrième ontologie pour modéliser les compétences, et étendu l'ontologie des tâches de diagnostic à des tâches quelconques. Ce logiciel a été réalisé dans le cadre d'un contrat de 3 ans avec la société performanSE SA, et a fait l'objet d'un transfert technologique ayant donné naissance au produit Atanor-knowesia (Athanor est la version recherche, la version industrielle est sans « h »).

1.6. Ontologies et Web sémantique

Le Web sémantique[32] s'attache à réintroduire du sens et donc de la connaissance sur le contenu de la Toile. Il offre un ensemble de techniques basées sur XML afin de représenter la sémantique dans des ontologies formelles et d'en permettre le traitement par des requêteurs et des raisonneurs.

Dans cette optique web sémantique, nous nous sommes intéressés à la modélisation de connaissances pour des agents émotionnels supportés par une plateforme multi-agents. A cette fin, nous avons proposé de coupler un modèle cognitif BDI (Belief, Desire Intension) à une modélisation agent UML. Cette formalisation permet de décrire les mécanismes d'évolution interne des agents, ainsi que leurs interactions, et débouche sur la production d'une base de connaissances en RDF/OWL[33]. Ce travail, dont l'objectif in fine est de produire un système d'aide à la décision sur la dynamique des groupes, a été impulsé par le projet ARTA sur l'étude du comportement en milieu professionnel des victimes d'un traumatisme crânien. Il s'est ensuite prolongé dans le cadre d'un projet RIAM : Groupe d'Agents Collaboratifs Emotionnels (GRACE).

Plus récemment, en nous inspirant de nos travaux sur les règles d'association, nous avons proposé une méthode, extensionnelle et asymétrique, originale pour l'alignement d'ontologies

---

[29] Système d'Aide à la MAintenace des Trieuses Automatiques

[30] J. Philippé, F. Guillet, D. Follut, P. Vandekerckhove. Un serveur de connaissances dans un contexte de maintenance appliquée aux machine de tri postal. *Journées Internationales Ingénierie des systèmes et NTIC, NimesTIC'2000*, pages 30-35, 2000.

[31] D. Follut, F. Guillet, J. Philippé, P. Vandekerckhove. ATHANOR - Un système pour la capitalisation et le déploiement de connaissances de diagnostic. Revue *Extraction des Connaissances et Apprentissage (ECA)*, vol. 1, n° 1-2, 315-324, 2001. Hermès.

[32] M. C. Daconta, L. J. Obrst, K. T. Smith (2003). *The Semantic Web : A Guide to the Future of XML, Web Services, and Knowledge Management*. Kluwer.

[33] S. Daviet, H. Desmier, H. Briand, F. Guillet, and V. Philippe. A System of Emotional Agents for Decision-Support. *IEEE/WIC/ACM Int. Conf. on Intelligent Agent Technology,IAT'05*, pp. 711-717, 2005, IEEE Press.

définies sous la forme de taxonomies instanciées sur un corpus textuel[34],[35]. L'idée directrice de notre approche est d'aider un utilisateur à aligner deux ontologies en lui proposant des règles d'association entre concepts. Plus précisément, nous considérons qu'une ontologie est un graphe orienté de concepts structurés par une relation de subsomption, où chaque concept est décrit dans des documents textuels contenant des termes caractéristiques. Notre approche se décompose alors en deux étapes consécutives fondées sur la mesure d'intensité d'implication : (1) l'extraction dans les documents des termes caractéristiques de chaque concept ; puis (2) l'extraction d'un ensemble minimal, au sens d'un critère de réduction des redondances, de règles d'association entre concepts.

1.7. Perspectives

En synthèse, les fondamentaux de mes travaux axés sur l'aide à la décision sont : le traitement des connaissances en tant que données complexes (mesures de qualité, ontologies, web sémantique), et les processus de fouille anthropocentrée (utilisateur, interactivité, visualisation).

J'envisage de prolonger de cette activité en la projetant dans la direction des deux verrous scientifiques majeurs souvent mentionnés en ECD :

- le passage à l'échelle sur les *masses de connaissances*,

- et la *fouille de connaissances* vues comme des données complexe.

De mon point de vue, cette perspective a aussi l'avantage d'allier des enjeux *académiques* et des enjeux *d'applications en entreprises*.

Je discerne une clé, sous-jacente à ces 2 verrous, cachée dans l'établissement de *ponts scientifiques* entre la *gestion/ingénierie des connaissances* et la *fouille de données*.

1.7.1. Passage à l'échelle sur les masses de connaissances

Sur les données très volumineuses, ou « masses de données », la plupart des approches classiques voient leurs performances s'effondrer dramatiquement. Ceci provient du fait que la complexité des algorithmes sous-jacents dépendent du nombre d'enregistrements (i.e. plus d'un million) ; et dépendent, souvent plus fortement encore, du nombre de variables traitées (i.e. plus d'un milliers). Deux types de solutions, une sur les données, l'autre sur les algorithmes, sont envisageables. La première, bien traitée dans la littérature, consiste à réduire la taille des données par échantillonnage des enregistrements et/ou par réduction des variables, mais elle incorpore un risque de perte d'information. La seconde, moins traitée, consiste à réduire la complexité des algorithmes soit en recherchant la linéarité, soit en distribuant les calculs sur une infrastructure virtuelle (grid computing ou grille de calcul), soit en quittant les approches globales pour passer à des approches locales incorporant des connaissances du domaine, ou des heuristiques d'exploration.

Cette seconde piste me semble la plus prometteuse. Nous avons commencé à l'aborder dans notre approche anthropocentrée de fouille de règles, où les connaissances de l'utilisateur dirigent implicitement, telle une heuristique, un algorithme de fouille local. J'envisage de

[34] J. David, F. Guillet, R. Gras, H. Briand. Alignement extensionnel et asymétrique d'ontologies par découverte d'implications entre concepts. *Revue Nationale des Technologies de l'Information (RNTI)*, E6, pp 151-162, 2006.

[35] J. David, F. Guillet, R. Gras, H. Briand. Conceptual hierarchies matching: an approach based on discovery of implication rules between concepts. In *17th European Conference on Artificial Intelligence (ECAI)*. Riva del Garda, Italy, Aug 28 - Sept 1, 2006. IOS Press. To appear.

prolonger cette approche, en créant un « pont scientifique » avec l'ingénierie des connaissances, qui permettra d'utiliser explicitement les connaissances de l'utilisateur après les avoir extraites et formalisée. Les travaux sur les mesures de qualité subjectives sont de cet ordre, mais ils se heurtent à la difficulté de l'extraction préalable des connaissances du domaine.

En aval du processus d'ECD, le passage à l'échelle se traduit aussi sur les « masses de connaissances » lorsque les algorithmes produisent de grandes quantités d'informations (comme les règles d'association). Comme précédemment, les approches locales restent intéressantes. Mais, dans ce contexte, j'envisage une autre perspective qui s'appuierait sur un couplage avec l'ingénierie des connaissances et plus précisément avec les grilles de connaissances. En effet, les grilles de connaissances (ou semantic grid[36]) constituent une approche récente, issue du croisement des grilles de calcul et des techniques du web sémantique, qui a l'avantage d'offrir une infrastructure collaborative de stockage et de services interopérables ouvrant l'accès au traitement de très grandes masses de connaissances. Ainsi, tirant bénéfice du support offert par ces grilles, je propose de développer un service d'annotation sémantique des connaissances, où l'utilisateur pourrait restituer une sémantique personnalisée aux connaissances intéressantes une fois celles-ci repérées dans la masse des résultats.

1.7.2. Fouille de connaissances

Stimulés par l'apparition et sur le web et dans les bases de données du langage XML, nouveau standard interopérable pour les données semi-structurées, les données complexes se banalisent. La nature complexe des données est liée à leur structure (graphes, dimensions spatiale et temporelles…), à leur hétérogénéité (textes, données, multimédia), et enfin à leur localisation (multi-sources).

Ayant toujours à l'esprit le pont évoqué entre l'ECD et la gestion des connaissances, je discerne trois perspectives autour de la fouille des données complexes. (1) Partant du constat que les connaissances peuvent êtres considérées comme des données à structure complexe, la première perspective consiste à développer des algorithmes spécifiques de fouille de données dans des bases de connaissances (composées d'ontologies OWL, de règles, …), afin de découvrir des relations (méta-connaissances ?) entre les connaissances. (2) En transposant les techniques de fouille pour l'analyse des traces d'utilisation dans les log des sites web, on peut s'intéresser au développement d'algorithmes de fouille spécifiques pour analyse les traces d'usage des connaissances dans les bases de connaissances. Cette perspective constitue une extension naturelle de mes travaux sur le serveur de connaissances Athanor. (3) Elle peut aussi être appliquées à l'analyse des traces d'activité laissées par les agents émotionnels décrits précédemment.

[36] M. Cannataro and D. Talia. Semantics and Knowledge Grids: Building the Next-Generation Grid. IEEE Intelligent Systems, pp. 56-63, Vol. 19, No 1, 2004.

2. PUBLICATIONS

2.1. Liste classée des publications

Livres

[1] H. Briand et F. Guillet (eds). Extraction et Gestion des Connaissances. Numéro spécial, *Revue Extraction des Connaissances et Apprentissage (ECA)*, vol. 1, n° 1-2, 2001. Hermès Science Publication. ISBN 2-7462-0216-6.

[2] H. Briand, M. Sebag, R. Gras, F. Guillet (eds). Mesures de qualité pour la fouille de données. Numéro spécial, *Revue Nationale des Technologies de l'Information (RNTI)*, E1, 2004, Cépaduès. ISBN 2.85428.646.4.

[3] F. Guillet and H. Hamilton (eds). Quality Measures in Data Mining. *Studies in Computational Intelligence*, 313 pages, april 2007, Springer. ISBN 3540449116. To appear.

[4] R. Gras, E. Suzuki, F. Guillet and F. Spagnolo (eds). Statistical Implicative Analysis. To appear in 2007.

Chapitres de livres

[5] J. Blanchard, P. Kuntz, F. Guillet, R. Gras. Implication Intensity: from the basic statistical definition to the entropic version, *Statistical Data Mining and Knowledge Discovery*, Chap. 8, 475-493, 2003. Chapman & Hall, CRC Press, ISBN 1584883448.

[6] X.-H. Huynh, F. Guillet, J. Blanchard, P. Kuntz, R. Gras and H. Briand. A graph-based clustering approach to evaluate interestingness measures : a tool and a comparative study. *Quality Measures in Data Mining*, *Studies in Computational Intelligence*, 2007. Springer ISBN 3540449116. To appear

Revues internationales

[7] J. Blanchard, F. Guillet and H. Briand. Interactive Visual Exploration of Association Rules with the Rule Focusing Methodology. *Knowledge and Information Systems (KAIS)*, 2006. Springer. ISSN: 0219-1377. To appear.

Revues nationales

[8] E. Pichon, P. Lenca, F. Guillet, et J. W. Wang. Un algorithme de partition d'un produit direct d'ordres totaux en un nombre minimum de chaînes. *Mathématiques Informatique et Sciences Humaines*, 32e année (125):5–15, 1994.

[9] R. Gras, P. Kuntz, R. Couturier, F. Guillet. Une version entropique de l'intensité d'implication pour les corpus volumineux. *Extraction des Connaissances et Apprentissage (ECA)*, vol. 1, n° 1-2, 69-80, 2001. Hermès Science Publication. ISBN 2-7462-0216-6.

[10] D. Follut, F. Guillet, J. Philippé, P. Vandekerckhove. ATHANOR - Un système pour la capitalisation et le déploiement de connaissances de diagnostic. *Extraction des Connaissances et Apprentissage (ECA)*, vol. 1, n° 1-2, 315-324, 2001. Hermès Science Publication. ISBN 2-7462-0216-6.

[11] P. Kuntz, F. Guillet,R. Lehn, H. Briand. Vers un processus d'extraction de règles d'association centré sur l'utilisateur. *In Cognito, Revue Francophone internationale en Sciences Cognitives*, n°20, p. 13-26, 2001. ISSN 1267-8015.

[12] T. Teusan, G. Nachouki, F. Guillet, H. Briand. La découverte de règles d'association dans les bases de données denses - Une approche génétique. *Extraction des Connaissances et Apprentissage (ECA)*, vol. 1, n° 4, 53-64, 2002. Hermès Science Publication. ISBN 2-7462-0506-1.

[13] J. Blanchard, F. Guillet, H. Briand. L'intensité d'implication entropique pour la recherche de règles de prédiction intéressantes dans les séquences de pannes d'ascenseurs. *Extraction des Connaissances et Apprentissage (ECA)*, vol. 1, n° 4, 77-88, 2002. Hermès Science Publication. ISBN 2-7462-0506-1.

[14] R. Gras, F. Guillet, R. Gras, J. Philippé. Réduction des colonnes d'un tableau de données par quasi-équivalence entre variables. *Extraction des Connaissances et Apprentissage (ECA)*, vol. 1, n° 4, 197-202, 2002. Hermès Science Publication. ISBN 2-7462-0506-1.

[15] J. Blanchard, F. Guillet, F. Rantière, H. Briand. Vers une Représentation Graphique en Réalité Virtuelle pour la Fouille Interactive de Règles d'Association. *Extraction des Connaissances et Apprentissage (ECA)*, vol. 17, n° 1-2-3, 105-118, 2003. Hermès Science Publication. ISSN 0992-499X, ISBN 2-7462-0631-5.

[16] J. Blanchard, F. Guillet, H. Briand. Une visualisation orientée qualité pour la fouille anthropocentrée de règles d'association. *In Cognito - Cahiers Romans de Sciences Cognitives*, n° 1.3, pp 79-100, 2004. ISSN 1267-8015.

[17] J. Blanchard, F. Guillet, R. Gras, H. Briand. Mesurer la qualité des règles et de leurs contraposées avec le taux informationnel TIC. *Revue Nationale des Technologies de l'Information (RNTI)*, E-2, vol. 1, pp 287-298, 2004. Cépaduès. ISBN 2.85428.633.2.

[18] J. Blanchard, P. Kuntz, F. Guillet, R. Gras. Mesure de qualité des règles d'association par l'intensité d'implication entropique. *Revue Nationale des Technologies de l'Information (RNTI)*, E1, pp 33-44, 2004, Cépaduès. ISBN 2.85428.646.4.

[19] R. Lehn, F. Guillet, H. Briand. Qualité d'un ensemble de règles : élimination des règles redondantes. *Revue Nationale des Technologies de l'Information (RNTI)*, E1, pp 141-168, 2004, Cépaduès. ISBN 2.85428.646.4.

[20] S. Daviet, F. Guillet, A. Magda Florea, H. Briand, V. Philippé. Modélisation des interactions entre individus avec AgentUML. *Revue Nationale des Technologies de l'Information (RNTI),* E3, pp 613-624, 2005, Cépaduès. ISBN 2.85428.677.4.

[21] H. Desmier, F. Guillet, A. Magda Florea, H. Briand, V. Philippé. Modélisation d'un agent émotionnel en UML et RDF. *Revue Nationale des Technologies de l'Information (RNTI)*, E3, pp 637-642, 2005, Cépaduès. ISBN 2.85428.677.4.

[22] J. Blanchard, F. Guillet, H. Briand, R. Gras. Une version discriminante de l'indice probabiliste d'écart à l'équilibre pour mesurer la qualité des règles. *Revue Quaderni di Ricerca in Didattica*, n°15, pp 131-138, 2005. ISSN 1592-5137.

[23] J. David, F. Guillet, V. Philippé, H. Briand, R. Gras. Validation d'une expertise textuelle par une méthode de classification basée sur l'intensité d'implication. *Revue Quaderni di Ricerca in Didattica*, n°15, pp 157-162, 2005. ISSN 1592-5137.

[24] X.-H. Huynh, F. Guillet, H. Briand. ARQAT : une plateforme d'analyse exploratoire pour la qualité des règles d'association. *Revue Quaderni di Ricerca in Didattica*, n°15, pp 339-349, 2005. ISSN 1592-5137.

[25] X.-H. Huynh, F. Guillet, H. Briand. Comparaison des mesures d'intérêt de règles d'association : une approche basée sur des graphes de corrélation. *Revue Nationale des Technologies de l'Information (RNTI)*, E6, pp 549-560, 2006, ISSN 1764-1667.

[26] R. Gras, J. David, J.-C. Régnier, F. Guillet. Typicalité et contribution des sujets et des variables supplémentaires en Analyse Statistique Implicative. *Revue Nationale des Technologies de l'Information (RNTI)*, E6, pp 359-370, 2006, ISSN 1764-1667.

[27] J. David, F. Guillet, R. Gras, H. Briand. Alignement extensionnel et asymétrique d'ontologies par découverte d'implications entre concepts. *Revue Nationale des Technologies de l'Information (RNTI)*, E6, pp 151-162, 2006, ISSN 1764-1667.

[28] B. Pinaud, P. Kuntz, F. Guillet, V. Philippé. Visualisation en Gestion des Connaissances : développement d'un nouveau modèle graphique Graph'Atanor. *Revue Nationale des Technologies de l'Information (RNTI)*, E6, pp 311-322, 2006, ISSN 1764-1667.

Communications invitées, tutoriels

[29] F. Guillet. Mesures de la qualité des connaissances en ECD. Actes des tutoriels, *4ème Conférence francophone Extraction et Gestion des Connaissances (EGC'2004)*, 120 pages, Clermont-Ferrand, Janvier 2004.

[30] F. Guillet, P. Lenca. Quality in Data Mining. *Invited session, 11th International Symposium on Applied Stochastic Models and Data Analysis, ASMDA'2005*, Brest, France, May 17-20, 2005. ISBN 2-908849-15-1.

Conférences internationales (avec comité de sélection)

[31] F. Guillet and G. Coppin. Industrial process control : experimental design and expertise acquisition. In *First European Conference on 'Cognitive Science in Industry'*, R. Bisdorff ed., pages 93–116. Luxembourg, September 1994.

[32] J.-P. Barthélemy, G. Coppin, and F. Guillet. Smelting process control : from experimental design to acquisition of expertise. In *International Conference on Industrial Engineering and Production Management*, vol. II, pages 2-11, Marrakech, March 1995.

[33] S. Guillaume, F. Guillet and J. Philippé. Contribution of the integration of intensity of implication into the algoritm proposed by Agrawal. In *14th European Meeting on Cybernetics and Systems Research, EMCSR'98*, volume 2, pages 805-810, Vienna, Austria, April 14-17 1998. Austrian Society of Cybernetic Studies. ISBN 3-85206-139-3.

[34] R. Lehn, F. Guillet and H. Briand. Eliminating redundant knowledge in an association rule-based system : an algorithm. In *14th European Meeting on Cybernetics and Systems Research, EMCSR'98*, volume 2, pages 793-798, Vienna, Austria, April 14-17 1998. Austrian Society of Cybernetic Studies. ISBN 3-85206-139-3.

[35] S. Guillaume, F. Guillet, and J. Philippé. Improving the discovery of association rules with intensity of implication. In *Second European Symp. on Principles of Data Mining and Knowledge Discovery, PKDD'98*. Principles of Data Mining and Knowledge Discovery, LNCS, vol. 1510, pages 318-327,

Nantes, France, 1998. Springer, ISBN 3-540-65068-7.

[36] F. Guillet, R. Lehn, P. Kuntz. A Genetic Algorithm for Visualizing Networks of Association Rules. In *12th Int. Conf. on Industrial & Engineering Applications of A. I. & Expert Systems, IEA/AIE-99*. Multiple Approaches to Intelligent Systems, LNCS, vol. 1611, pages 145-154, Cairo, May 31- June 3 1999. Springer. ISBN 3-540-66076-3.

[37] R. Lehn, F. Guillet, P. Kuntz, H. Briand and J. Philippé. Felix : An interactive rule mining interface in a kdd process. In *10th Mini-Euro Conference, Human Centered Processes, HCP'99*, pages 169-174, Brest, France, September 22-24, 1999.

[38] P. Kuntz F. Guillet, R. Lehn and H. Briand. A User-Driven Process for Mining Association Rules. In *4th European Conf. on Principles of Data Mining and Knowledge Discovery, PKDD'2000. Principles of Data Mining and Knowledge Discovery*, LNCS, vol. 1910, pp 483-489, Lyon, France, September 12-16, 2000. Springer. ISBN 3-540-41066-X.

[39] J. Blanchard, F. Guillet, H. Briand. A Virtuel Reality Environment for Knowledge Mining. In *11th Mini-Euro Conference, Human Centered Processes, HCP'2003: Distributed Decision Making and Man-Machine Cooperation*, pp 175-179, Luxembourg, May 5-7, 2003.

[40] J. Blanchard, F. Poulet, F. Guillet, P. Kuntz. Highly Interactive Data Mining with Virtual Reality. In *5th Virtual reality International Conference, VRIC'2003*, pp 221-228, Laval, May 14-16, 2003, ISBN 2-9515730-2-2.

[41] J. Blanchard, F. Guillet, H. Briand. A User-driven and Quality oriented Visualization for Mining Association Rules. In *3rd IEEE International Conference on Data Mining, ICDM'2003*, pp 493-497, Melbourne, Florida, USA, November 19 - 22, 2003, IEEE Computer Society Press.

[42] R. Couturier, R. Gras, F. Guillet. Reducing the number of variables using implicative analysis. In *Int. Federation of Classification Societies, IFCS'2004*, pp 277-285, Chicago, USA, July 15-18, 2004. Springer Verlag : Classification, Clustering, and Data Mining Applications. ISBN 3-540-22014-3.

[43] J. Blanchard, F. Guillet, H. Briand, R. Gras. Assessing the interestingness of rules with a probabilistic measure of deviation from equilibrium. In *11th Int. Symposium on Applied Stochastic Models and Data Analysis, ASMDA'2005*, J. Janssen and P. Lenca (eds), pp 191-199, Brest, France, May 17-20, 2005. ISBN 2-908849-15-1

[44] J. David, F. Guillet, V. Philippé, R. Gras. Implicative statistical analysis applied to clustering of terms taken from a psychological text corpus. In *11th Int. Symposium on Applied Stochastic Models and Data Analysis, ASMDA'2005*, J. Janssen and P. Lenca (eds), pp 201-208, Brest, France, May 17-20, 2005. ISBN 2-908849-15-1

[45] X.-H. Huynh, F. Guillet, H. Briand. ARQAT: An Exploratory Analysis Tool For Interestingness Measures. In *11th Int. Symposium on Applied Stochastic Models and Data Analysis, ASMDA'2005*, J. Janssen and P. Lenca (eds), pp 334-344, Brest, France, May 17-20, 2005. ISBN 2-908849-15-1

[46] X.-H. Huynh, F. Guillet, H. Briand. Clustering Interestingness Measures with Positive Correlation. In *7th Int. Conf. on Enterprise Information Systems, ICEIS'2005*, pp 248-253, Miami, USA, May 24-28, 2005, INSTICC. ISBN 972-8865-19-8.

[47] S. Daviet, H. Desmier, H. Briand, F. Guillet, and V. Philippe. A System of Emotional Agents for Decision-Support. In *IEEE/WIC/ACM Int. Conf. on Intelligent Agent Technology,IAT'05*, pp. 711-

717, 2005, IEEE Computer Society Press. ISBN 0-7695-2416-8.

[48] X.-H. Huynh, F. Guillet, H. Briand. A data analysis approach for evaluating the behavior of interestingness measures. In *8th Int. Conf. on Discovery Science, DS'05*, LNCS 3735 XVI, pp 330 – 337, 2005, Springer. ISBN: 3-540-29230-6.

[49] J. Blanchard, F. Guillet, R. Gras, H. Briand. Using Information-theoretic Measures to Assess Association Rule Interestingness. In *5th IEEE Int. Conf. on Data Mining, ICDM'2005*, pp. 66-73, 2005, IEEE Computer Society Press.

[50] X.-H. Huynh, F. Guillet, H. Briand. Extracting representative measures for the post-processing of association rules. In *Fourth IEEE Int. Conf. on Computer Sciences dedicated to Research, Innovation and Vision for the Future (RIVF'06)*, pp 99-106, 2006, IEEE Computer Society Press.

[51] X.-H. Huynh, F. Guillet, H. Briand. Discovering the Stable Clusters between Interestingness Measures. In *8th Int. Conf. on Enterprise Information Systems, ICEIS'2006*, pp 196-201, 2006. ISBN 972-8865-41-4.

[52] X.-H. Huynh, F. Guillet, H. Briand. Evaluating interestingness measures with correlation graph. In *19th Int. Conf. on Industrial, Engineering & Other Applications of Applied Intelligent Systems, IEA/AIE'06*, lncs 4031, pp 312 - 321, 2006. Springer. ISBN 3-540-35453-0.

[53] X.-H. Huynh, F. Guillet, H. Briand. A graph-based approach for comparing interestingness measures. In *First IEEE International Conference on Engineering of Intelligent Systems (IEEE ICEIS'06), pp 375-380*. Islamabad, Pakistan. 2006. IEEE Computer Society Press.

[54] J. David, F. Guillet, R. Gras, H. Briand. Conceptual hierarchies matching: an approach based on discovery of implication rules between concepts. In *17th European Conference on Artificial Intelligence (ECAI)*, pp 357-361, 2006. IOS Press. ISBN 1-58603-642-4.

[55] B. Pinaud, P. Kuntz, F. Guillet, V. Philippé. Graphical Visualization in the Knowledge Management System Atanor. In *6th International Conference on Knowledge Management (I-KNOW'06)*, pp 481-488, 2006. *Journal of Universal Computer Science (J.UCS)*, Springer. ISSN 0948-6968.

[56] J. David, F. Guillet, H. Briand. Mapping directories and OWL ontologies with AROMA. In *ACM Conference on Information and Knowledge Management (CIKM)*, Arlington, USA, november 2006. ACM. To appear.

Ateliers internationaux (avec comité de sélection)

[57] F. Guillet, D. Follut, P. Van De Kerckhove and J. Philippé. Samanta: Towards Using Virtual Reality in an Computer-Assisted Environment for the Maintenance of Postal Sorting Machines. In J. Tisseau and G. Subsol (Eds), In *first French-British International Workshop on Virtual Reality*, page 135, July 11-12, 2000, Brest, France.

[58] F. Guillet, P. Vandekerckhove. Samanta : a Knowledge Server. In J.-L. Ermine (editor), *Workshop on Knowledge Management - Theory and Application in conjunction with PKDD'2000*, pp 19-28, Lyon, France, September 12, 2000.

[59] F. Guillet. A Human-centered Process for Data Mining. In *European Operational Research conference (EURO'2001)*, p. 93, Rotterdam, The Netherlands, july 9-11, 2001

[60] J. Blanchard, F. Guillet, H. Briand. Exploratory Visualization for Association Rule Rummaging. In *4th International Workshop on Multimedia Data Mining, MDM/KDD2003, in conjunction with KDD'2003*, pp 107-114, Washington DC, USA, August 24-27, 2003.

[61] J. David, F. Guillet, R. Gras, H. Briand. An interactive, asymmetric and extensional method for matching conceptual hierarchies. In *Open INTEROP Workshop On "Enterprise Modelling and Ontologies for Interoperability" (INTEROP-EMOI06)*. Luxembourg, June 5-6, 2006. To appear.

Conférences nationales (avec comité de sélection)

[62] S. Guillaume, F. Guillet, et J. Philippé. L'intensité d'implication pour la découverte de règles d'association ''pertinentes''. Actes des *1ères journées Ré-ingéniérie des Systèmes d'Information (RSI'98)*, pages 36-43, Lyon, France, 1-2 Avril 1998.

[63] S. Guillaume et F. Guillet. Une généralisation des règles d'association par l'intensité d'implication ordinale. Actes des *7èmes Journées de la Société Francophone de Classification*, pages 77-86, Nancy, France, Septembre 1999.

[64] R. Lehn, P. Kuntz, F. Guillet. Sur un problème de tracé de graphes posé par un processus interactif d'extraction de règles d'associations. Actes du *3ème congrès de la société Francaise de Recherche Opérationnelle et d'Aide à la Décision, ROADEF'2000*, pages 151-156, Nantes, France, 26-28 Janvier 2000.

[65] J. Philippé, F. Guillet, D. Follut, P. Vandekerckhove. Un serveur de connaissances dans un contexte de maintenance appliquée aux machine de tri postal. *Journées Internationales Ingénierie des systèmes et NTIC, NimesTIC'2000*, pages 30-35, Nîmes, France, 11-13 Septembre 2000.

[66] R. Lehn, F. Guillet, P. Kuntz, H. Briand, J. Philippé. Félix : un outil interactif d'aide à la fouille de connaissances s'appuyant sur l'intensité d'implication. R. Gras et M. Bailleul (eds), *Fouille dans les données par la méthode d'analyse statistique implicative*, pages 51-58. Association pour la recherche en Didactique des Mathématiques. Septembre 2000. ISBN 2-9516505-0-7.

[67] D. Follut, F. Galletti, F. Guillet. Les Arbres de Décision Interactifs (ADI). Actes du *14ème Forum National de la Maintenance*, Paris, Novembre 2002, éditions AFIM.

[68] F. Guillet, D. Follut, V. Philippé, J. Philippé. ATHANOR : Une approche pour la gestion des connaissances de maintenance sur des systèmes complexes. *Première journée d'Etude sur les Systèmes d'Information pour l'Aide à la Décision en Ingénierie Système, JESIADIS'2002*, pages 41-54, ENSIETA, Brest, France, 28 novembre 2002.

Conférences et ateliers nationaux à publication restreinte (avec comité de sélection)

[69] F. Guillet, P. Lenca et E. Pichon. Extraction de règles de décision en situation d'expertise. *Premier Colloque Jeunes Chercheurs en Sciences Cognitives*, Grenoble, mars 1994.

[70] R. Gras, F. Guillet, P. Peter et J. Philippé. Apprentissage automatique et implication : mise en oeuvre sur un espace d'apprentissage en ressources humaines. Actes de la *quatrièmes journées de la Société Francophone de Classification*, Vannes, 19-20 Septembre 1996.

[71] R. Lehn, F. Guillet, P. Kuntz. Félix : un outil interactif d'aide à la fouille de connaissances s'appuyant sur l'intensité d'implication. Actes des *journées "fouille dans les données par la méthode d'analyse statistique implicative"*, Caen, 23-24 Juin 2000.

[72] J. Blanchard, R. Lehn, F. Guillet, P. Kuntz. Des graphes à la réalité virtuelle pour l'extraction adaptative de règles d'association. Actes de l'atelier *Visualisation et Extraction Adaptative des Connaissances*, conférence EGC2003, pages 24-25, Lyon, 22 janvier 2003.

[73] X.-H. Huynh, F. Guillet, H. Briand. ARQAT : une plateforme d'analyse exploratoire pour la qualité des règles d'association. Actes de l'atelier *Qualité des Données et des Connaissances (QDC2005)*, conférence EGC2005, pp 58-68, 18 janvier 2005, Paris.

[74] J. Blanchard, F. Guillet, H. Briand, R. Gras. IPEE : Indice Probabiliste d'Ecart à l'Equilibre pour l'évaluation de la qualité des règles. Actes de l'atelier *Qualité des Données et des Connaissances (QDC2005)*, conférence EGC2005, pp 26-34, 18 janvier 2005, Paris.

[75] R. Gras, R. Couturier, F. Guillet, F. Spagnolo. Extraction de règles en incertain par la méthode implicative. Actes de l'atelier *Qualité des Données et des Connaissances (QDC2005), conférence EGC2005*, pp 19-25, 18 janvier 2005, Paris.

[76] J. David, F. Guillet, V. Philippé, H. Briand, R. Gras. Validation d'une expertise textuelle par une méthode de classification basée sur l'intensité d'implication. Actes de *l'atelier Qualité des Données et des Connaissances (QDC2005)*, conférence EGC2005, pp 50-57, 18 janvier 2005, Paris.

[77] N. Ronarc'h, G. Rozec, F. Guillet, A. Nédélec, S. Baquedano, V. Philippé. Modélisation des connaissances émotionnelles par les cartes cognitives floues. Actes de *l'atelier Modélisation des Connaissances*, conférence EGC2005, pp 11-20, 18 janvier 2005, Paris.

[78] M. Ahlonsou'h, E. Blanchard, H. Briand, F. Guillet. Transcription des concepts du diagramme de classe UML en OWL Full. Actes de l'atelier *Modélisation des Connaissances*, conférence EGC2005, pp 63-76, 18 janvier 2005, Paris.

[79] C. Marinica, Y. Fossé, S. Daviet, H. Briand, F. Guillet. Représentation d'expertise psychologique sous la forme de graphes orientés, codés en RDF. Actes de la *conférence EGC2006*, pp 713-714, 17 janvier 2006, Lille, Poster.

[80] X.-H. Huynh, F. Guillet, H. Briand. Extraction de mesures d'intérêt représentatives pour le post-traitement des règles d'association. Actes de l'atelier *Qualité des Données et des Connaissances (QDC2006)*, conférence EGC2006, pp 45-54, 17 janvier 2006, Lille.

[81] J. David, F. Guillet, R. Gras, H. Briand Alignement de taxonomies documentaires : une méthode asymétrique et extensionnelle. Actes de la conférence *Ingénierie des Connaissances (IC2006)*, Mai 2006, Nantes. Poster. A paraître.

[82] C. Marinica, F. Guillet, H. Briand Représentation d'interactions entre agents par des règles RIS formalisées en RDF. Actes de la conférence *Ingénierie des Connaissances (IC2006)*, Mai 2006, Nantes. Poster. A paraître.

[83] S. Daviet, F. Guillet, H. Briand, V. Philippé La simulation multi-agents pour l'aide à la décision. Actes de la conférence *Ingénierie des Connaissances (IC2006)*, Mai 2006, Nantes. Poster. A paraître.

Rapports

[84] J. Blanchard, P. Kuntz, F. Guillet, R. Gras. Mesure de la qualité des règles d'association par l'intensité d'implication entropique. AS GafoDonnées - groupe GafoQualité. Novembre 2002, CNRS.

[85] R. Gras, R. Couturier, M. Bernadet, J. Blanchard, H. Briand, P. Kuntz, F. Guillet, R. Lehn, P. Peter. Un exemple de mesure de Qualité : l'implication statistique. AS GafoDonnées - groupe GafoQualité. Novembre 2002, CNRS.

[86] F. Guillet. Rapport d'activité et projets de recherche sur la qualité des connaissances. AS GafoDonnées. Paris V Descartes, 5 novembre 2002, CNRS.

[87] R. Gras et F. Guillet (eds). Actes de la 1ère journée Qualité des Connaissances (GafoQualité). Polytech'Nantes, 21 Mars 2002.

Autres : séminaires, …

[88] F. Guillet. Contributions à la maîtrise de qualité d'un processus industriel par apprentissage des stratégies de contrôle de l'opérateur. *Thèse de Doctorat*, Université de Rennes I, Décembre 1995.

[89] F. Guillet. Extraction automatique de règles de décision. Journées du PRC Discrimination Symbolique/Numérique, Saint-Malo, Octobre 1993.

[90] F. Guillet. Partition d'un produit direct d'ordres totaux en un nombre minimum de chaînes. Journées Mathématiques Discrètes et Sciences Sociales. Brest, Juin 1993.

[91] F. Guillet. Conférencier Invité. Table ronde "Mémoire d'entreprise, quels enjeux?". Salon Formatec'98. CNIT, Paris La Défense, 1-4 Décembre 1998.

[92] F. Guillet, H. Briand. Sensibilisation à la problématique de mesure de la qualité des connaissances. Journée Modélisation et Extraction des Connaissances, CPER COM. Angers/Laval/Nantes. Nantes, 6 Mai 1999.

[93] H. Briand, F. Guillet. Indices de mesure de la qualité des connaissances. Journée Modélisation et Extraction des Connaissances, CPER COM. Angers/Laval/Nantes. Angers, 16 Décembre 1999.

[94] R. Lehn, F. Guillet. Felix: un environnement graphique de fouille de connaissances. Journée Modélisation et Extraction des Connaissances, CPER COM. Angers/Laval/Nantes. Laval, 6 Avril 2000.

[95] F. Guillet. Mesures de la qualité des connaissances : mesures globales versus locales. Journée Modélisation et Extraction des Connaissances, CPER COM. Angers/Laval/Nantes. Nantes, 15 Mars 2001.

[96] F. Guillet. Mesure de qualité par l'intensité d'implication. Journée Modélisation et Extraction des Connaissances, CPER COM. Angers/Laval/Nantes. Laval, 7 Février 2002.

[97] F. Guillet, R. Gras, P. Kuntz, H. Briand. Mesure de qualité de règles d'association par analyse implicative. R. Gras et F. Guillet (eds), actes de la 1ère journée Qualité des Connaissances (GafoQualité), page 5. Polytech'Nantes, 21 Mars 2002.

[98] F. Guillet. Bilan et perspectives de recherche sur la qualité des connaissances. 3ère journée AS GafoDonnées - STIC - CNRS. IHP, Paris, 28 Mars 2002.

[99] F. Guillet. Qualité des connaissances : synthèse. 2ème journée GafoQualité - AS GafoDonnées - STIC - CNRS. Polytech'Nantes, 20 Juin 2002.

[100] J. Blanchard, F. Guillet, H. Briand. Représentation en Réalité Virtuelle pour la Validation de Règles d'Associations. 2ème Journée Modélisation et Extraction des Connaissances, CPER COM.

Angers/Laval/Nantes. Angers, 8 avril 2003.

[101] F. Guillet. Evaluation de la qualité des connaissances. 1ère journée du groupe 3.4 "Fouille de Données" du PRC-GDR I3. Paris V Descartes, 12 juin 2003.

[102] J. David, F. Guillet. Typicalité et contributions de variables supplémentaires en analyse statistique implicative. Journée Modélisation et Extraction des Connaissances, CPER COM. Angers/Laval/Nantes. Laval, 16 décembre 2004.

[103] J. David, F. Guillet. Validation d'une expertise textuelle par une méthode de classification basée sur l'intensité d'implication. Journée Modélisation et Extraction des Connaissances, CPER COM. Angers/Laval/Nantes. Nantes, 17 février 2005.

[104] C. Marinica, F. Guillet. Représentation d'expertise psychologique sous la forme de graphes orientés, codés en RDF. Journée Modélisation et Extraction des Connaissances, CPER COM. Angers/Laval/Nantes. Nantes, 5 janvier 2006.

[105] F. Guillet. Fouille de trace d'agents émotionnels dans un système multi-agents. Journée GRACE, RIAM. Brest/Nantes. Nantes, 15 mars 2006.

[106] F. Guillet. Fouille de données, enjeux et perspectives. *Journée Atlanticiel*, CCI Nantes, 29 juin 2006.

2.2. Sélection de Publications

Les publications sélectionnées pour les rapporteurs visent à illustrer différentes contributions aux problématiques présentées dans l'introduction de ce chapitre :

- Mesures de qualité :

F. Guillet and H. Hamilton (eds). Quality Measures in Data Mining. *Studies in Computational Intelligence*. 2006. Springer. To appear.

- Fouille anthropocentrée :

J. Blanchard, F. Guillet and H. Briand. Interactive Visual Exploration of Association Rules with the Rule Focusing Methodology. *Knowledge and Information Systems (KAIS)*, 2006. Springer. ISSN: 0219-1377. To appear

J. Blanchard, F. Guillet, H. Briand. A User-driven and Quality oriented Visualization for Mining Association Rules. In *3rd IEEE International Conference on Data Mining, ICDM'2003*, pp 493-497, 2003, IEEE Press.

- Serveur de connaissances et ontologies :

D. Follut, F. Guillet, J. Philippé, P. Vandekerckhove. ATHANOR - Un système pour la capitalisation et le déploiement de connaissances de diagnostic. Revue *Extraction des Connaissances et Apprentissage (ECA)*, vol. 1, n° 1-2, 315-324, 2001. Hermès.

- Ontologies et Web sémantique :

S. Daviet, H. Desmier, H. Briand, F. Guillet, and V. Philippe. A System of Emotional Agents for Decision-Support. In *IEEE/WIC/ACM Int. Conf. on Intelligent Agent Technology,IAT'05*,

pp. 711-717, 2005, IEEE Press.

J. David, F. Guillet, R. Gras, H. Briand. Conceptual hierarchies matching: an approach based on discovery of implication rules between concepts. In *17th European Conference on Artificial Intelligence (ECAI)*. Riva del Garda, Italy, Aug 28 - Sept 1, 2006. IOS Press. To appear.

3. RESPONSABILITES EN RECHERCHE

3.1. Responsabilité au sein d'équipes de recherche

3.1.1. Co-responsable du thème "Extraction et Gestion des Connaissances" [2002-2004]

Pendant deux ans, j'ai animé avec Pascale Kuntz le thème Extraction et Gestion des Connaissances (EGC) de l'équipe Langues et Connaissances (LEC) de l'Institut de Recherche en Informatique de Nantes (IRIN). Ce thème regroupe 8 enseignants chercheurs permanents répartis sur trois sites nantais (Polytech'Nantes, IUT OGP, Formation continue de l'Université), 4 chercheurs associés et 7 doctorants, dont 4 en co-encadrement avec d'autres institutions (Ensieta, Brest et Esiea, Laval)

Les activités de recherche du thème EGC s'articulent autour de quatre trois axes principaux : (1) l'évaluation de la qualité des connaissances et les algorithmes de fouille ; (2) les aides à la visualisation pour l'extraction adaptative ; (3) la gestion et l'ingénierie des connaissances.

3.1.2. Co-Responsable de deux thèmes de l'équipe COD [1999-…]

Depuis 1999, je co-anime deux axes de recherche axes qui sont devenus des thèmes de l'équipe COD du LINA en 2004. Le premier thème « Mesure de la qualité des connaissances » regroupe aujourd'hui 4 permanents et 3 doctorants. Le second thème « Mesure de la qualité des connaissances » regroupe aujourd'hui 4 permanents et 1 doctorant.

Historiquement, peu après mon arrivée à Nantes en 1997, j'ai contribué activement avec Henri Briand et Jacques Philippé, puis avec Pascale Kuntz à partir de 1999, au renforcement et à l'animation des deux thèmes « mesures de qualité » et « gestion des connaissances » de l'équipe CID de l'IRIN. Ces thématiques de recherche ont perduré au fil des réorganisations. En effet, en 2002, lors d'une première réorganisation de l'IRIN, l'équipe CID est devenue le thème EGC, que j'ai co-animé avec Pascale Kuntz de 2002 à 2004, au sein de l'équipe fédératrice LEC dirigée par Henri Briand. Enfin en 2004, lors de la création du laboratoire FRE CNRS LINA, les équipes fédératrices ont disparues, et le thème EGC est devenu l'équipe COD. Ma co-responsabilité du thème EGC s'est arrêtée avec cette dernière, lors de laquelle co-animation d'équipe n'était pas prévue dans les statuts du nouveau laboratoire.

3.2. Chargé de Communication de l'association de recherche EGC [2002-…]

Après avoir organisé la première édition de la conférence EGC à Nantes en 2001, j'ai oeuvré à la fondation de l'association « Extraction et Gestion des Connaissances » (EGC, association loi 1901, SIRET 444 256 507 00013).

Puis en 2003, j'ai été élu à la fonction de *chargé de communication* du bureau de l'association (2 mandats de 2 ans depuis 2003).

L'association EGC a pour vocation d'animer la recherche dans les domaines de l'extraction des connaissances dans les données et de la gestion des connaissances, et de renforcer les liens avec les entreprises. Elle supporte la conférence annuelle EGC, et participe depuis 2005 aux Rencontres Inter-Association (RIA) avec les sociétés connexes (AFIA, ARIA, ASTI, INFORSID, SFC, SFDS, SPECIF, LMO).

Le bureau de l'association est composé de : H. Briand (président), D. Zighed (vice-président), D. Hérin (secrétaire), R. Gras (trésorier), et F. Guillet (chargé de communication).

Ma fonction de *Chargé de communication* concerne : la coordination inter-conférences, la gestion des parrainages et des relations avec les sociétés scientifiques connexes, la gestion des adhérents, la responsabilité logistique (conception et mise en œuvre) du site web (voir http://www.polytech.univ-nantes.fr/associationEGC/) et de la liste de diffusion des annonces (liste-egc@polytech.univ-nantes.fr, comportant plus de 1000 adhérents francophones).

Parallèlement, depuis 2003, je suis membre permanent du comité de pilotage d'EGC. Ce comité de pilotage composé de 12 membres permanents est destiné à pérenniser le cycle de conférences EGC. Pour cela, il est amené à choisir : les organisateurs des éditions d'EGC à partir des dossiers de candidature déposés, les présidents des comités d'organisation et de programme, les présidents d'honneur, les conférenciers invités, et des lauréats pour la remise des prix.

J'ai organisé, en tant que président du comité d'organisation, la première édition de la conférence EGC à Nantes en 2001. Voici un récapitulatif de l'ensemble des éditions d'EGC :

- 2001, à Nantes, présidents H. Briand (CP) et F. Guillet (CO), 4 tutoriels, 120 participants.

- 2002, à Montpellier, présidents D. Zighed (CP) et D. Hérin (CP et CO), 6 tutoriels, 150 participants.

- 2003, à Lyon, présidents M.-S. HACID (CP), Y. Kodratoff (CP) et D. Boulanger (CO), 6 tutoriels, 200 participants.

- 2004, à Clermont-Ferrand, présidents G. Piatetsky-Shapiro (PH), G. Hébrail (CP), L. Lebart (CP), J.-M. Petit (CO), 2 tutoriels, 6 ateliers, 250 participants.

- 2005, à Paris, présidents G. Piatetsky-Shapiro (PH), N. Vincent (CP et CO), S. Pinson (CP), 2 tutoriels, 12 ateliers, 300 participants.

- 2006, à Lille, présidents H. Mannila (PH), G. Ritschard(CP), C. Djeraba (CO), 2 tutoriels, 9 ateliers, 250 participants.

- 2007, à Namur, Belgique, présidents E. Suzuki (PH), G. Venturini (CP), M . Noirhomme (CO).

3.3. Coordination internationale du Master ECD et du Master européen MDM&KD [2002-…]

Depuis 2002, et en complémentarité de ma responsabilité pédagogique des relations internationales du département informatique de Polytech'Nantes (voir Partie IV :1.1 Responsable des Relations Internationales, page 65), je suis chargé de la coordination internationale de deux Masters.

3.3.1. Coordination internationale du Master ECD [2002-…]

Depuis 2002, je suis chargé de la coordination internationale du Master "Extraction de Connaissances dans les Données" (Master ECD). Ce master de recherche (M2) dirigé par Lyon 2 (D. Zighed) est co-habilité par les universités de Lyon 2, Nantes, et Paris 11. Il a la particularité d'offrir ses enseignements en téléenseignement par visioconférence sur 4 sites (Lyon, Nantes Paris, et Bucarest) et d'accueillir des étudiants internationaux de Roumanie (Politehnica Bucarest) et du Vietnam (University of Cantho).

Après avoir initié et développé les collaborations avec la Roumanie et le Vietnam, ma fonction consiste à gérer la coordination avec ces deux universités, selon 2 activités principales :

- la responsabilité administrative et pédagogique des échanges internationaux d'étudiants en stage ou en semestre d'étude, et d'enseignants invités (AUF, SOCRATES, EGIDE),
- La gestion du système de téléenseignement par visioconférence réparti sur les 4 sites (financement, création, déploiement et gestion).

3.3.2. Création et coordination internationale du Master européen MDM&KD [2006-…]

Depuis 2006, je participe à l'élaboration et au lancement du projet de Master européen « Master in Data Mining and Knowledge Discovery » (MDM&KD). Ce nouveau master constitue une extension européenne et anglophone du Master ECD. Il débutera en septembre 2006 et, en plus des partenaires Roumains et Vietnamiens, intégrera un troisième partenaire Italien : l'université du Piémont Oriental.

Je suis chargé d'établir et de coordonner la liaison avec ce troisième partenaire, selon les mêmes modalités celles établies pour le Vietnam et la Roumanie (échanges internationaux, et déploiement du système de visioconférence).

A partir de septembre 2007, mon activité devra s'étendre à deux nouveaux partenaires : l'université de Liège en Belgique et l'université de Dortmund en Allemagne.

3.4. Responsabilités éditoriales

Sous l'impulsion d'Henri Briand, je me suis impliqué dans l'éditions de quatre ouvrages avec comité de lecture : deux numéros spéciaux de revues nationales, et deux livres de chapitres internationaux. Ces ouvrages sont intimement reliés à mes thématiques de recherche en fouille de données. J'ai réalisé l'intégralité du processus d'édition de ces quatre ouvrages : la constitution du comité de lecture, le site sur la toile des appels, le processus d'évaluation, la mise en forme finale, les négociations avec l'organisme de publication…

3.4.1. Ouvrages Internationaux

Il s'agit de deux recueils de chapitres de visibilité internationale comportant un comité de lecture composé en majorité de chercheurs internationaux dans le domaine du Data Mining. Chaque chapitre a fait l'objet de trois évaluations réalisées par au moins 2 membres du comité de lecture : une évaluation de propositions sur résumé, puis une évaluation du chapitre soumis, et enfin une évaluation du chapitre révisé. Le premier ouvrage est finalisé et sera publié par Springer en 2006. Le second est en cours de construction et sa publication en prévue en 2007.

1. F. Guillet and H. Hamilton (eds). *Quality Measures in Data Mining*. Studies in Computational Intelligence. 2007. Springer. ISBN 3540449116. To appear

Ce livre de chapitre s'inscrit dans le prolongement international des activités de recherche du groupe GafoQualité (AS GafoDonnées 2002-2003). Les membres du comité de lecture, parrainé par *Gregory Piatetsky-Shapiro (USA, ACM SIGKDD chair)*, sont :

Henri Briand (France), Régis Gras (France), Yves Kodratoff (France), Vipin Kumar (USA), Pascale Kuntz (France), Robert Hilderman (Canada), Ludovic Lebart (France), Philippe Lenca (France), Bing Liu (USA), Amédéo Napoli (France), Gregory Piatetsky-Shapiro (USA), Gilbert Ritschard (Switzerland), Sigal Sahar (USA), Gilbert Saporta (France), Dan Simovici (USA), Jaideep Srivastava (USA), Einoshin Suzuki (Japan), Pang-Ning Tan (USA), Alexander Tuzhilin (USA), Djamel Zighed (France).

2. R. Gras, E. Suzuki, F. Guillet, and F. Spagnolo (eds). *Statistical Implicative Analysis: Theory and applications*. To appear in 2007.

Ce deuxième ouvrage est orienté vers l'usage de l'analyse statistique implicative, notamment développé dans le cadre du cycle de conférences ASI. Il cible les éléments statistiques et les applications de cette technique. La parution est prévue pour 2007.

3.4.2. Ouvrages Nationaux

Il s'agit de deux numéros spéciaux de revues nationales avec comité de lecture.

1. H. Briand et F. Guillet (eds). Extraction et Gestion des Connaissances. Numéro spécial, *Revue Extraction des Connaissances et Apprentissage (ECA)*, vol. 1, n° 1-2, 2001. Hermès Science Publication. ISBN 2-7462-0216-6.

Ce numéro spécial est issu de la conférence EGC01 que j'ai organisée à Nantes en janvier 2001.

2. H. Briand, M. Sebag, R. Gras, F. Guillet (eds). Mesures de qualité pour la fouille de données. Numéro spécial, *Revue Nationale des Technologies de l'Information (RNTI)*, E1, 2004, Cépaduès. ISBN 2.85428.646.4.

Ce numéro spécial dresse un état de l'art synthétisant les activités du groupe GafoQualité de l'Action Spécifique GafoDonnées que j'ai animé en 2002 et 2003 (voir Partie III :3.4, page 42).

3.5. Co-animateur du groupe GafoQualité - AS GafoDonnées - STIC CNRS. [2002-2003]

Au sein de l'Action Spécifique GafoDonnées (AS STIC CNRS), comportant 29 équipes et 70 membres, animée par R. Cicchetti et M. Sebag, j'ai co-animé avec R. Gras le *groupe de travail GafoQualité*. Pendant 2 années le groupe GafoQualité a rassemblé 30 chercheurs répartis dans 10 équipes nationales. Il a permis d'identifier et de dynamiser les recherches nationales sur le thème des « mesures de qualité en fouille de données », et son activité a été récapitulée dans un rapport CNRS. Il a ensuite débouché sur la rédaction d'un numéro spécial « Qualité des connaissances » de la revue RNTI paru en 2004. Depuis 2005, ce groupe poursuit son activité dans le cadre de l'association EGC, et a animé 2 ateliers « Qualité des Données et des Connaissances » dans les conférences EGC'05, et EGC'06.

3.6. Responsabilités sur des contrats de recherche

- Co-responsable avec Henri Briand pour l'équipe COD de l'axe MEC-EGC du projet état-région CPER COM (Connaissances Objets Modélisation). Animation et production des rapports d'activité [2002-2004]. Voir Partie III :5.2, page 49.
- Co-responsable avec Henri Briand pour l'équipe COD du projet RIAM GRACE. Gestion administrative et encadrement de 4 CDD [2004-2006]. Voir Partie III :5.3, page 49.
- Co-responsable avec Jacques Philippé des contrats industriels Samanta [1996-2000], Atanor [1999-2004]. Voir Partie I :1, page 53.

3.7. Autres responsabilités en recherche

- Membre élu de la commission de spécialistes, section 27. [2001-2004, puis réélu depuis 2004]

4. ENCADREMENTS

4.1. Thèses

Thèse de Rémi Lehn - Convention CIFRE - Soutenue le 15 décembre 2000 à l'Université de Nantes - *Titre : Un système interactif de visualisation et de fouille de règles pour l'extraction de connaissances dans les bases de données*. L'objectif de la thèse menée en collaboration avec la société PerformanSE SA (Carquefou, 44) était de contribuer au développement d'un logiciel implémentant un processus de fouille de règles d'association centré sur l'utilisateur. Création d'un prototype : FELIX. Participation à l'encadrement : 50%. R. Lehn est actuellement maître de conférences à l'Université de Nantes.

Thèse de Sylvie Guillaume - Contrat avec LA POSTE- le 15 décembre 2000 à l'Université de Nantes - *Titre : Traitement des données volumineuses - Mesures et algorithmes d'extraction de règles d'association et de règles ordinales*. L'objectif de la thèse menée en sous-contrat avec LA POSTE (Chartres, 28) était de construire des mesures de qualité adaptées à la quantification de tendances implicatives entre variables numériques. Participation à l'encadrement : 50%. S. Guillaume est actuellement maître de conférences à l'Université de Clermont-Ferrand.

Thèse de Julien Blanchard – Bourse de la fondation VédiorBis, Fondation de France - Soutenue le 24 novembre 2005 à l'Université de Nantes - *Titre : Un système de visualisation pour l'extraction, l'évaluation, et l'exploration interactives des règles d'association.* L'objectif de la thèse était de développer un processus de fouille de règles d'association centré sur l'utilisateur et fondé sur des métaphores en réalité virtuelle et des mesures de qualité. Création d'un prototype : ARVIS. Participation à l'encadrement : 80%. J. Blanchard est actuellement maître de conférences à Polytech'Nantes.

Thèse de Xuan-Hiep Huynh – Bourse d'excellence du gouvernement Vietnamien - Débutée en septembre 2003, soutenance prévue en décembre 2006 à l'Université de Nantes - *Titre : Evaluation des mesures d'intérêt pour le post-traitement des règles d'association en ECD.* Dans une perspective d'analyse exploratoire des données, l'objectif de la thèse est de recenser, d'étudier et de comparer la panoplie des mesures de qualité de règles (mesures d'intérêt) disponibles. Création d'un prototype : ARQAT. Participation à l'encadrement : 70%.

Thèse de Jérôme David – Bourse de la fondation VédiorBis, Fondation de France - Débutée en janvier 2004, soutenance prévue en 2007 à l'Université de Nantes - *Titre : Aide à l'alignement d'ontologie par une méthode extensionnelle, orientée et interactive*. L'objectif de la thèse est de proposer une approche interactive d'aide à l'alignement d'ontologies instanciées sur des documents, en utilisant des techniques de fouille de texte (extraction terminologique) et de fouille de données (règles d'association). Création d'un prototype en cours. Participation à l'encadrement : 70%.

Thèse de Stéphane Daviet - Convention CIFRE - Débutée en décembre 2004, soutenance prévue en décembre 2007 à l'Université de Nantes -*Titre : Modélisation et extraction de connaissances sur un simulateur de dynamique de groupe*. L'objectif de la thèse menée en collaboration avec la société Knowesia (Carquefou, 44) est de formaliser le comportement et les connaissances d'un groupe d'agents émotionnels simulant le comportement d'un groupe humain, puis d'en extraire les comportements émergents par des techniques de fouille de données. Création d'un prototype en cours. Participation à l'encadrement : 80%.

Thèse de Claudia Marinica – Bourse régionale - Débutée en octobre 2006, soutenance prévue en 2009 à l'Université de Nantes -*Titre : Fouille interactive de connaissances à l'aide de mesures d'intérêt et de représentations sémantiques*. L'objectif de la thèse est d'incorporer des mécanismes de gestion des connaissances dans un processus de fouille de règles d'association afin d'améliorer la représentation des connaissances du domaine et des préférences du décideur afin d'améliorer l'extraction des meilleures règles. Participation à l'encadrement : 80%.

4.2. DEA et MASTER

1. P. Merheb (1999). Visualisation de règles d'associations. DEA informatique, université de Nantes.
2. C. Archaux (2000). Visualisation de connaissances en réalité virtuelle pour l'ECD. DEA informatique, université de Nantes.
3. J. Clech (2000). Métaphores 3D de visualisation de règles en ECD. DEA informatique, université de Nantes.
4. T. Teusan (2000). La découverte rapide de règles d'associations. DEA informatique, université de Nantes.
5. J. Blanchard (2001). Découverte de connaissances dans des séquences de pannes d'ascenseurs. DEA ECD, Polytech'Nantes.
6. J. Laumonier (2002). Coupling Knowledge Management and Deductive Database systems. DEA ECD, Polytech'Nantes. Co-encadrement avec I. Vlahavas, Aristotle University of Thessaloniki, Grèce.
7. F. Rantière (2002). Représentation de connaissances en réalité virtuelle pour l'extraction de connaissances dans les données. DEA ECD, Polytech'Nantes.
8. E. Popovici (2003). Un atelier pour l'évaluation des indices de qualité. DEA ECD, Polytech'Nantes. Co-encarement avec S. Traussan-Matu, Politehnica Bucarest, Roumanie.
9. G. De Renty (2003). Visualisation de règles d'associations. DEA informatique, université de Nantes. Co-encadrement avec F. Poulet, ESIEA Recherche Laval.
10. C. Gerzé (2003). Formalisation et intégration d'une expertise SIG sur le cabotage dans un outil d'aide à la décision. DEA ECD, Polytech'Nantes.
11. C. Nedu (2004). A local approach for extracting association rules. DEA ECD, Polytech'Nantes. Co-encadrement avec B. Duval, LERIA, Angers.
12. H. Desmier (2004). Modélisation d'un agent émotionnel avec Agent-UML. DEA ECD, Polytech'Nantes. Co-encadrement avec A. Magda Florea, Politehnica Bucarest, Roumanie.
13. S. Daviet (2004). Modélisation et extraction de connaissances sur une plateforme multi-agents. DEA ECD, Polytech'Nantes. Co-encadrement avec S. Traussan-Matu, Politehnica Bucarest, Roumanie.
14. L. Georgescu (2005). Modélisation BDI des connaissances d'un agent émotionnel dans Jadex. Master ECD, Polytech'Nantes. Co-encadrement avec A. Magda Florea, Politehnica Bucarest, Roumanie.
15. L. Georget (2005). Extraction de connaissances spatio-temporelles. Master ECD, Polytech'Nantes.

16. H. El Attar (2005). Analyse d'un questionnaire par des techniques d'ECD. Master ECD, Polytech'Nantes.
17. M. Albert (2005). Modélisation d'un agent cognitif en RDF. Master ECD, Polytech'Nantes. Co-encadrement avec A. Magda Florea, Politehnica Bucarest, Roumanie.
18. C. Marinica (2006). Modélisation de règles d'interactions par approche web sémantique. Master ECD, Polytech'Nantes.

4.3. Jurys de thèses Externes

1. Co-rapporteur de la thèse de H. Cherfi (2004). Etude et réalisation d'un système d'extraction de connaissances à partir de textes. Université Henri Poincaré - Nancy 1.
2. Co-rapporteur de la thèse de D. Tanasa (2005). Web Usage Mining: Contributions to Intersites Logs Processing and Sequential Pattern Extraction with Low Support. Université de Nice Sophia Antipolis.

5. RECHERCHE CONTRACTUELLE

5.1. Réseau d'excellence –Interop NoE (PCEU)

Titre du projet : Projet européen « Interoperability Research for Networked Enterprises Applications and Software » du 6ème Programme Cadre UE - Interop NoE

Partenaires : Interop regroupe 47 partenaires sur 15 pays, dont l'équipe COD du LINA.

Durée : 3 ans (2003-2006) - Montant global 12 M€.

Objectifs : Ce projet Européen vise à créer les conditions pour mener des recherches compétitives et innovantes dans le domaine de l'interopérabilité des applications et des logiciels d'entreprises. Ce projet est animé par M ; Harzallah au sein de l'équipe COD, en liaison avec G. Berio de l'université de Turin.

Contribution : je participe avec M. Harzallah à 3 workpackages :

- WP1 : définition de la carte des compétences et connaissances des participants au projet Interop,
- WP5 : définition d'un modèle unifié de représentation de l'entreprise, et étude des langages et outils pour son implémentation,
- WP8 : Etudes des ontologies et définition d'une ontologie générique pour l'entreprise.

5.2. Contrat de Plan Etat Région - Pôle Informatique Régional (STIC 8).

Titre du projet : Projet COM - Axe Modélisation et Extraction de Connaissances (MEC)

Partenaires : L'axe MEC regroupe 3 laboratoires (LINA, Nantes - LERIA, Angers - ESIEA Recherche, Laval)

Durée : 4 ans (1999 - 2006) - Montant équipe : 10 k€ par an.

Objectifs (pour l'équipe CID) : Ce contrat permet de constituer un axe régional de recherche autour de la visualisation adaptative pour l'extraction de connaissances, et de développer des contacts pour les recherches sur l'évaluation de la qualité des connaissances.

Contribution : ma contribution a permis de renforcer les liens entre l'équipe COD et les deux autres équipes régionales : avec l'ESIEA sur la visualisation (réunions communes semestrielles, 1 stage de Master commun, 1 communication commune à VRIC'03, Laval), et avec le LERIA sur les mesures de qualité (1 stage de Master commun, un numéro spécial dans la revue RNTI E1, 2004 en synergie avec le groupe GafoQuallité de l'AS GafoDonnées). J'ai également été chargé de la rédaction du rapport d'activité de l'axe durant 3 ans de 2002 à 2004.

5.3. Réseau pour la Recherche et l'Innovation en Audiovisuel et Multimedia (RIAM).

Titre du projet : Groupe d'Agents Collaboratifs Emotionnels (GRACE).

Partenaires : projet piloté par J. Tisseau (ENIB) regroupant des membres de 2 laboratoires (Lab. Informatique Industrielle, ENIB, Brest - LINA, Nantes) et de 2 entreprises (PerformanSE SA, Nantes et Virtualys, Brest)

Durée : 24 mois (Mars 2004 à Mars 2006), Montant équipe : 74 k€.

Objectifs : Le projet a trait à l'animation comportementale et l'autonomie d'entités intelligentes. Il vise à concevoir un modèle original destiné à une application ludique qui permet d'intégrer dans une simulation participative multi-agents des interactions basées sur des échanges relationnels entre agents et utilisateurs d'un environnement virtuel ayant divers profils comportementaux émotionnels.

Contribution : Co-responsable scientifique avec H. Briand pour l'équipe COD – LINA, j'ai assuré la gestion de 4 contrats CDD, et animé une équipe de 10 personnes (3 CNAM, 3 Master, et 4 projets ingénieur). Notre contribution scientifique a été double. Elle a porté d'une part sur la modélisation de l'expertise psychologique avec l'utilisation conjointe d'AgentUML, d'un modèle BDI et d'OWL. Et d'autre part sur l'utilisation des techniques de fouille de données spatio-temporelles afin d'analyser les résultats des simulations.

5.4. Fondation Recherche et Emploi VediorBis - Fondation de France.

Titre du projet : Aides à la visualisation pour la fouille de règles

Partenaires : projet piloté par P. Kuntz- en collaboration avec la sociéte VediorBis

Durée : 2004-2006, Montant équipe : 20 k€.

Objectifs : Ce projet vise à développer des outils visuels adaptés à la manipulation dynamique d'ensembles de règles d'association. L'attention portera en particulier sur la structuration de ces ensembles en entités entretenant entre elles des relations implicatives significatives selon des critères de qualité statistiques. Une instanciation sera faite sur un corpus de curriculum vitae prétraité par des outils de text-mining pour trouver des tendances implicatives entre des facteurs déterminants (typologies comportementales, ...).

Contribution : ma contribution scientifique a porté sur le développement d'une approche graphique et interactive, basée sur des métaphore en réalité virtuelle pour la fouille de règles d'association.

5.5. Fondation Recherche et Emploi VediorBis - Fondation de France.

Titre du projet : INférences sur un modèle de gestion Couplée des Compétences et des Connaissances (INF3C)

Partenaires : projet piloté par M.Harzallah, en collaboration avec Politehnica Bucarest et l'université de Turin.

Durée : 2003 - 2005 - Montant équipe : 20 k€

Objectifs : Le projet pour objectif de proposer un système intelligent de gestion intégrée des compétences et des connaissances qui doit permettre : *(i)* le raisonnement sur la représentation définie en vue d'un enrichissement dynamique, *(ii)* une réponse aisée à différentes requêtes, notamment la détermination de la liste des compétences par acteur associées à une typologie prédéfinie et de la liste des compétences associées à un curriculum vitae à partir de la liste des connaissances spécifiées pour un poste.

Contribution : ma contribution a porté sur la partie représentation des connaissances et des compétences.

5.6. Projet avec le Service de Maintenance des Installations de LA POSTE.

Titre du projet : Système d'Aide à la Maintenance des Trieuses Automatiques (SAMANTA)

Partenaires : projet co-piloté par J. Philippé et F. Guillet, en collaboration avec LA POSTE,

SMIP, Chartres.

Durée : 1998 - 2001 - Montant équipe : 100 k€ (financement d'une thèse)

Objectifs : Le projet pour objectif de proposer un système intelligent pour la gestion des connaissances de diagnostic de pannes sur les machines de tri automatique de courrier postal. Dans une perspective de gestion des connaissances, l'approche choisi a consisté à s'inspirer du modèle de diagnostic de KADS, pour produire une approche reposant sur (i) 3 vues ontologiques (tâches, composants, symptômes), associées à (ii) un puissant éditeur graphique, et à (iii) des représentations des machines en réalité virtuelle, et enfin (iv) embarquant un raisonneur permettant l'aide au diagnostic.

Contribution : j'ai été le principal maître d'oeuvre de ce projet, et ai conçu le logiciel SAMANTA, qui implémente cette approche.

5.7. Projet avec PerformanSE SA.

Titre du projet : Serveur de connaissances ATHANOR

Partenaires : projet co-piloté par J. Philippé et F. Guillet, en collaboration avec la société PerformanSE SA, Carquefou.

Durée : 2001 - 2004 - Montant équipe : 40 k€

Objectifs : Le projet pour objectif la conception d'un « serveur de connaissances » muni de bonnes performances de représentation, d'évolutivité, et de déploiement, afin de répondre aux besoins de l'entreprise en matière de gestion des connaissances. Ce projet a permis de consolider et d'améliorer et de généraliser à des tâches autres que le diagnostic les recherches effectuées dans le cadre du projet SAMANTA. En particulier, en termes de recherche, une quatrième ontologie portant sur les compétences à été conçue, et l'ontologie des tâches totalement repensée.

Contribution : j'ai été le responsable scientifique de ce projet, ai conçu un prototype nommé ATHANOR, et enfin j'ai assuré le transfert technologique avec la société PerformanSE SA. Ce projet a aboutit à un logiciel professionnel : ATANOR (sans h !) vendu par sa filiale Knowesia.

6. TRANSFERTS TECHNOLOGIQUES VERS LES ENTREPRISES

6.1. SAMANTA [1998-2001]

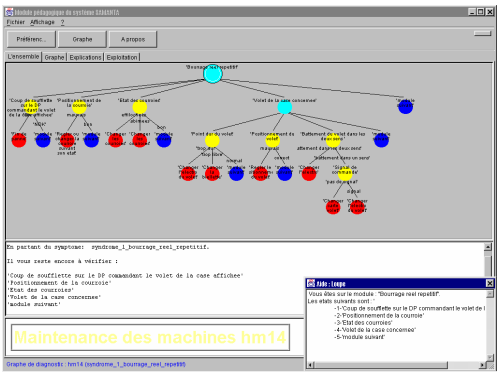
SAMANTA (Système d'Aide à la Maintenance des Trieuses Automatiques) est un logiciel de *gestion des connaissances* qui a été développé dans le cadre d'un contrat de 4 ans coordonné par J. Philippé avec le Service de Maintenance des Installation de LA POSTE (SMIP à Chartres). Il implémente les concepts issus de nos travaux de recherche, et a fait l'objet d'un transfert technologique vers LA POSTE.

La phase de livraison et de déploiement a été effectuée en 2001. L'outil a préalablement été déployé sur 2 centres de tri pilotes : à Paris et à Nantes. Puis, après 6 mois de test l'outil a été déployé en 2001 sur les 50 centres nationaux de tri postal.

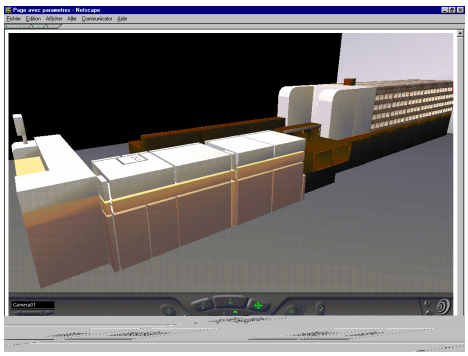
J'ai été le concepteur et le maître d'œuvre de ce projet de grande dimension et de longue durée, puisqu'il a nécessité 4 ans de développement et d'interactions avec La Poste. J'ai géré une équipe de projet constituée d'étudiants en stages pour un total de 70 homme-mois. La plus grande partie a été réalisée par des étudiants ingénieurs en stages et projets : 6 projets de 300 heures, 4 projets de 150 heures, 2 stages DUTIL de 5 mois, 4 stages ingénieur de 5 mois, soit un total de 46 hmois. J'ai aussi réalisé un important travail de gestion de projet, de recueil d'expertise, de développements complémentaires et d'intégration estimé à 24 hmois.

D'un point de vue technologique, ce logiciel est structuré selon une architecture multi-tiers comportant un serveur prolog (raisonneur et base de connaissances), un serveur web (interfaces), et utilise les technologies de l'Internet pour ses interfaces : html, java, vrml. Il embarque les connaissances sous la forme de 3 modèles ontologiques issus de nos travaux de recherche : (i) une taxonomique des symptômes et des tâches, (ii) un logigramme de tâches de diagnostic, (iii) une décomposition organique de la machine couplée aux documentations techniques et à des représentations en réalité virtuelle de cette machine. Le logiciel offre de 2 fonctionnalités principales : (i) un mode expert à travers une éditeur graphique des modèles ontologiques, (ii) un mode praticien pour la résolution de pannes (inférence sur les connaissances : des symptômes aux pannes).

Parallèlement à la conception de l'outil final, qui à été conçu incrémentalement à partir des retours obtenus sur trois prototypes intermédiaires, j'ai mené une campagne de recueil de connaissances auprès de 3 experts de La Poste (10 entretiens de 2 h), qui nous a permis d'alimenter la base de connaissances de SAMANTA. Une modélisation complète en réalité virtuelle a été réalisée pour deux gammes de machines de tri : HM14-15 et TOP.



Représentation du diagnostic par « logigramme »



Machine de tri virtuelle modélisée en VRML

6.2. ATHANOR [2001-2004]

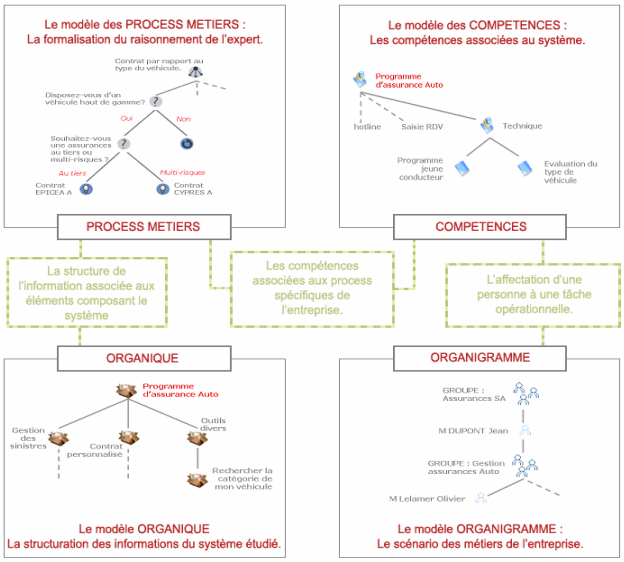
ATHANOR (creuset de l'alchimiste) est un *serveur de connaissances* qui a été développé dans le cadre d'un contrat de 4 ans avec la société PerformanSE SA (Carquefou). Ce logiciel s'est inscrit dans le prolongement du projet de gestion de connaissances mené avec LA POSTE. Ce logiciel a fait l'objet d'un important transfert technologique vers cette entreprise.

Le transfert technologique relatif au logiciel Athanor a été un succès. Une première version d'Athanor est sortie fin 2004, afin d'être proposée à l'essai auprès de grands comptes. Le marché ayant répondu favorablement, une édition professionnelle nommée ATANOR (sans h) est aujourd'hui commercialisée par la filiale Knowesia du groupe PerformanSE..

En collaboration avec Vincent Philippé (chef de projet à PerformanSE SA) qui a graduellement repris ce projet avec une équipe de 3 personnes, j'ai géré une équipe de projet pour un effort total de 60 homme-mois. La plus grande partie a été réalisée par des étudiants ingénieurs en stages et projets : 8 projets de 300 heures, 6 projets de 150 heures, 1 stages DUTIL de 5 mois, 4 stages ingénieur de 5 mois, soit un total de 47 hmois. Ma charge personnelle a été de 12 hmois pour la gestion du projet, l'intégration, et le transfert technologique.

D'un point de vue recherche, nous avons généralisé les modèles ontologiques conçus pour SAMANTA, qui étaient restreints aux tâches de diagnostic, en les étendant à la représentation de tâches quelconques. Nous avons aussi ajouté un quatrième modèle dédié à la gestion des compétences.

D'un point de vue technologique, l'architecture du logiciel a subi des évolutions majeures de ces interfaces et de ces serveurs, et a été rendue modulaire. De nombreuses fonctionnalités ont aussi été ajoutées afin de prendre en compte la gestion des compétences dans les modes expert (édition) et praticien (consultation). Au global le raisonneur du serveur de connaissances permet (en mode praticien) de répondre graphiquement aux questions : (i) *quoi* : quel est le problème, (ii) *comment* : comment le résoudre, (iii) *qui* : qui sait le faire ou de quelles compétences faut-il disposer.



Athanor : Les quatres modèles ontologiques et leur interactions

7. LOGICIELS

7.1. ARVAL [2003-2004]

ARVAL (Association Rules VALidation) est un prototype issu de nos travaux de recherche sur les mesures de qualité en fouille de données et développé dans le cadre du mémoire CNAM de L. Jaffrelot en 2004. Plus précisément, il s'agit d'un atelier permettant de calculer et les valeurs d'une trentaine de mesures d'intérêts sur un ensemble de règles d'association fourni en entrée. D'une manière surprenante, aucun outil n'était disponible pour cet usage, et nous avons décidé de combler cette lacune. D'un point de vue technique, l'outil est développé en java et s'appuie sur un formats de données normalisés, PMML, un dialecte XML dédié au data mining, afin de faciliter les échanges avec les autres outils d'ECD (Weka, SAS, …).

Ce logiciel freeware est téléchargeable à : http ://www.polytech.univ-nantes.fr/arval/

7.2. ARVIS [2002-2005]

ARVIS (Association Rules VISualization) est un logiciel issu de nos travaux de recherche (cf thèse de J. Blanchard, page 46). Il implémente notre approche d'aide à la décision pour la fouille interactive et locale de règles d'association.

L'outil client/serveur permet la connexion à un base de données contenant les données à traiter. Il implémente en interne notre algorithme de calcul local de règles d'association, et en sortie permet la visualisation interactive, dans un navigateur web, des sous ensembles de règles ciblés sous la forme de représentations graphiques en 3D (vrml). L'utilisateur dispose alors de 5 opérateurs d'interaction pour modifier les règles ciblées, et ainsi explorer interactivement l'ensemble des règles qui l'intéressent. La restitution peut être rendue immersive (réalité virtuelle), via l'utilisation d'un dispositif matériel stéréoscopique ad hoc (cf. salle réalité virtuelle, page 9)

Ce logiciel est aujourd'hui utilisé pour la validation d'une expertise psychologique.

7.3. ARQAT [2003-2006]

ARQAT (Association Rules Quality Analysis Tool) est un prototype issu de nos travaux de recherche (cf thèse de X.-H. Huynh, page 45). Il implémente notre approche d'analyse exploratoire des mesures d'intérêt de règles d'association à l'aide de graphes de corrélation.

L'outil permet de calculer les valeurs de 40 mesures d'intérêt sur un ensemble de règles d'association des règles, données en entrée. Puis, après application éventuelle d'un opération de sélection et d'échantillonnage, il offre une série d'outils d'analyse : (i) les résultats statistiques élémentaires incluant les cas limites, (ii) la visualisation des distributions croisées des mesures, (iii) l'analyse corrélative des mesure, (iv) la visualisation par des graphes de corrélation, (v) la comparaison sur une famille de ensembles de règles.

Ce logiciel est réalisé en java avec une standardisation XML/PMML des données. Il est entièrement graphique, et génère un rapport d'analyse dans un navigateur web. ARQAT est téléchargeable à : http ://www.polytech.univ-nantes.fr/arqat/

8. RAYONNEMENT SCIENTIFIQUE

8.1. Responsabilités scientifiques

8.1.1. Internationales

- *Editeur invité* d'un ouvrage "Quality Measures in Data Mining", Springer, A paraître en 2007.
- Membre du *réseau d'excellence* européen Interop-NoE. Participation à 3 workpackages [depuis 2003].

8.1.2. Nationales

- *Coordinateur* du cycle de conférences Extraction et Gestion des Connaissances [2001-2003].
- Membre *fondateur* et membre du *bureau* de l'association Extraction et Gestion des Connaissances [2002-…].
- Comité de *pilotage* : conférence Extraction et Gestion des Connaissances (EGC) [2003-…].
- Co-animateur responsable avec R. Gras du groupe de travail GafoQualité (30 chercheurs, 10 équipes) de l'Action spécifique GafoDonnée - STIC - CNRS. [2002-2003].
- Co-responsable avec H. Briand du projet RIAM GRACE : Groupe d'Agents Collaboratifs Emotionnels [2004-2006].
- Membre du projet AVIF : Aides Visuelles pour la Fouille de Règles [2004-2006].
- Membre du projet INF3C : INférences sur un modèle de gestion Couplée des Compétences et des Connaissances [2003-2005].

8.2. Comités de lecture de revues

8.2.1. Internationales

- Relecteur de European Journal of Operational Research, EJOR - Feature issue : Human Centered Processes, en **1998**, et **2004**.
- Relecteur de journal of Computational Statistics and Data Analysis, CSDA, **2006**.

8.2.2. Nationales

- Editeur invité du numéro spécial « Extraction et Gestion des Connaissances », Revue Extraction des Connaissances et Apprentissage (ECA), Hermès **2001**.
- Editeur invité du numéro spécial « Mesures de Qualité pour la Fouille de Données », Revue des Nouvelles Technologies de l'Information (RNTI), Cépaduès, **2004**.
- Relecteur Revue Ingénierie des Systèmes d'Information ISI, **2002**.
- Comité de lecture du numéro spécial « Modélisation des Connaissances », Revue des Nouvelles Technologies de l'Information (RNTI), Cépaduès, à paraître.

8.3. Comités de programme de conférences

8.3.1. Internationales

- IEEE Int. Conf. on Data Mining, 6th ICDM, Hong Kong, China **2006**. (ICDM est la conférence internationale phare en data mining)
- Int. Conf. on Discovery Science, DS, Padova Italy **2004**, puis Singapore **2005**, puis Barcelona Spain **2006**.
- IEEE Int. Conf. on Tools with Artificial Intelligence, 18th IEEE ICTAI, Washington D.C., USA, **2006**.
- Knowledge Acquisition Workshop, 9th Pacific Rim International Conference on Artificial Intelligence, PKAW/PRICAI, Guilin China **2006**.
- Int. conf. on Applied Stockastic Models for Data Analysis, ASMDA, Brest **2005**, puis Chania Crete Greece **2007**.
- Int. Workshop on Visual Data Mining, VDM/IEEE ICDM, Melbourne Florida USA en **2003**, puis à Brighton UK en **2004**.
- Int. conf. on Human-centred Process, HCP, Luxemburg **2003**, Vilnius Lithuania **2007**.
- Relecteur int. conf. on Principles of Knowledge Discovery in Data PKDD, Lyon **2000**, puis Freiburg Germany **2001**.
- Relecteur Int. Symp. on Methodologies for Intelligent Systems, ISMIS, Lyon **2002**, puis Maebashi Japan **2003**.
- Relecteur Int. Conf. on Advances in Information Systems, ADVIS, Izmir Turkey **2002**.
- Relecteur Int. conf. on Human-centred Process, HCP, Brest **1999**.
- Relecteur Int. Workshop on Knowledge Management - Theory and Application, KMTA/PKDD, Lyon **2000**.

8.3.2. Nationales

- *Président CP* de 2 ateliers «Qualité des Données et des Connaissances », QDC/EGC, Paris **2005**, Lille **2006**.
- Conférence Francophone Extraction et Gestion des Connaissances, EGC, Nantes **2001**, Montpellier **2002**, Lyon **2003**, Clermont-Ferrand **2004**, Paris **2005**, Lille **2006**, Namur **2007**
- Conférence Fouille dans les données par la méthode d'Analyse Statistique Implicative, ASI, Palerme **2005**.
- Relecteur INFORSID, Nantes **2002**, Biarritz **2004**.
- Relecteur ASI, Caen **2000**, Sao Paulo **2003**.
- Atelier Visualisation et Extraction de Connaissances, VEC/EGC, Lyon **2003**, Clermont-Ferrand **2004**.
- Atelier Modélisation et Extraction de Connaissances, MEC/EGC, Paris **2005**, Lille **2006**, Namur **2007**.

8.4. Organisation de conférences

- *Président* du *comité d'organisation* de la 1ère conférence francophone Extraction et Gestion des Connaissances, EGC, Nantes **2001**.
- Membre CO du 10ème Congrès INFORSID, Nantes, **2002**.
- Membre CO du Workshop on Knowledge Management - Theory and Application, KMTA/PKDD, Lyon **2000**.
- Atelier «Qualité des Données et des Connaissances », QDC/EGC, Paris **2005**, Lille **2006**, Namur **2007**.

- Journées GafoQualité, Nantes **2002**.
- Membre CO de la Semaine de la Connaissances, SdC, Nantes **2006**.

8.5. Conférences invitées et présidences de session

- Session invitée « Knowledge Quality », ASMDA, Brest **2005**.
- Tutoriel invité « Qualité des Connaissances », EGC Clermont-Ferrand **2004**.
- Présidence de session dans des conférences : HCP 1999, ASMDA **2005**.
- Présidence de session dans des conférences : EGC **2002**, **2003**, **2004**, **2005**, **2006**.

8.6. Coopérations internationales

8.6.1. Invitation de chercheurs étrangers

J'ai invité et accueilli 7 chercheurs d'universités étrangères :

- **G. Piatetski-Shapiro** (professeur, Chair of *ACM SIGKDD*, Chair of *KD Nuggets*, *USA*), que j'ai invité à être conférencier invité à EGC 2005, et qui parraine le comité de lecture de l'ouvrage QMDM, Springer 2006.
- **E. Suzuki** (professeur, University of Yokohama, Japan), chercheur invité 1 semaine en 2005 à Polytech'Nantes.
- **J. Srivastava** (professeur, Department of Informatics, University of Minnesota, USA), chercheur invité 1 mois en 2004 à Polytech'Nantes.
- **S. Simoff** (Professeur, University of Technology, Sydney, Austrialia), chercheur invité 1 mois en 2003 à Polytech'Nantes.
- **A. Magda-Florea** (Professeur, Politehnica, Bucarest, Romania), chercheur invité 1 mois en 2004 à Polytech'Nantes.
- **S. Traussan-Matu** (Professeur, Politehnica, Bucarest, Romania), chercheur invité 1 mois en 2003 à Polytech'Nantes.
- **E. Kalisz** (Professeur, Politehnica, Bucarest, Romania), chercheur invité 1 mois en 2000, puis 15 jours en 2006 à Polytech'Nantes.

8.6.2. Ecole Polytechnique de Bucarest (Universitatea Politehnica Bucuristi), Roumanie.

- *Contacts* : E. Andronescu (Présidente), G. Marian (Vice-Président), D. Popescu (Doyen), E. Kalisz (professeur, resp. Erasmus/Socratès), A. Magda Florea (professeur), S. Traussan-Matu (professeur, Académie des sciences, RACAI).

- *Type de collaboration* : La collaboration très forte en recherche depuis 2001 avec cette école de 30000 étudiants, classée première école Roumaine en infiormatique,. repose sur un master multi-sites commun, le Master ECD. Mon rôle sans cette collaboration s'est traduit par la mise en place du système de visio-conférences inter-sites à Polytech'Nantes, l'invitations de chercheurs (5 chercheurs invités 1 mois à Polytech'Nantes entre 2002 et 2005), des stages de masters co-encadrés avec bourses Socrates (7 stages de Master de 5 mois entre 2003 et 2005), et la visite en février 2006 à Polytech'Nantes des président, vice-président et doyen de cette école.

8.6.3. University of Cantho, Vietnam

- *Contacts* : T. Le Quyet (Doyen), G. Richard (professeur IFI Hanoï), T.-V. HO (professeur).

- *Type de collaboration* : La collaboration avec l'université de Cantho, a été initiée en 2003, avec la thèse en cotutelle de X.-H. Huynh (voir page 45). Dans le cadre du Master ECD ; et de financements de mobilité recherche de l'AUF, cette collaboration s'est renforcée à travers l'échange d'étudiants de master (1 en 2005, 2 en 2006). Un système de visio-conference est en cours de déploiement à Cantho. Mon rôle à consister à mettre en œuvre cette collaboration.

8.6.4. University of Piémont Oriental, Italy.

- *Contacts* : L. Saitta (professeur).

- *Type de collaboration* : cette collaboration a débuté en 2006, à travers le projet d'extension européenne du Master ECD, et sera amenée à se renforcer à partir de septembre 2006. Mon rôle consiste à coordonner cette collaboration : établissement des conventions d'échange interuniversitaire, gestion les supports d'échange Erasmus/Socrates, et mise en place du système de visio-conférences inter-sites.

8.6.5. University of Regina, Canada

- *Contacts* : H. Hamilton (professeur)

- *Type de collaboration* : depuis 2005, nous collaborons à la co-édition d'un ouvrage collectif de travaux de recherche, intitulé « Quality Measures in Data Mining », à paraître chez Springer en 2006.

8.6.6. University of Laval, Canada

- *Contacts* : N. Belkhiter (professeur)

- *Type de collaboration* : depuis 2005, nous avons conclu un accord d'échange entre nos deux universités, collaborons à travers l'échange de semestres étudiant en informatique niveau Master (5 semestres étudiants entre 2004 et 2006). Mon rôle a consisté à établir la convention d'échange, puis à gérer les échanges (stages, semestres, thèses).

8.6.7. University of Yokohama and Kyushu, Japan

- *Contacts* : E. Suzuki (professeur)

- *Type de collaboration* : E. Suzuki a été chercheur invité pour 1 mois à Polytech'Nantes en 1998. Depuis notre collaboration s'est renforcée à travers un ouvrage collectif (QMDM, Springer, 2006), et du projet en cours de co-éditer un livre intitulé « Implicative Statistical Analysis » en 2007. E. Suzuki a aussi participé comme conférencier invité à l'atelier QDC/EGC que j'ai co-organisé à Lille en 2006.

8.6.8. University of Technology, Sydney, Austrialia

- *Contacts* : S. Simoff (professeur)

- *Type de collaboration* : S. Simoff a été chercheur invité pour 1 mois à Polytech'Nantes en 2003. Nous avons ensuite collaboré à travers le workshop Visual Data Mining, VDM/IEEE ICDM dont j'ai été membre du comité de programme.

8.6.9. University of Palermo, Italy

- *Contacts* : F. Spagnolo (professeur)

- *Type de collaboration* : Nous avons co-rédigé un article (QDC 2005), et nous travaillons à la co-édition d'un livre intitulé « Implicative Statistical Analysis » prévu pour 2007.

8.6.10. Autres collaborations internationales

- *Aristotle University of Thessaloniki, Greece :* I. Vlahavas (professeur, Department of Computer Science and Engineering).

- Contacts individuels avec (1) H. Bock (*Aachen University, Germany*), (2) A. Freitas (*University of Kent, USA*), (3) R. Hilderman (*University of Regina, Canada*), (4) V. Kumar (*University of Minnesota, USA*), (5) B. Liu (*University of Illinois at Chicago (UIC), USA*), (6) G. Ritschard (*Université de Genève*), (7) S. Sahar (*Intel, USA*), (8) D. A. Simovici (*University of Massachusetts at Boston, USA*), (9) P.-N. Tan (*Michigan State University, USA*), (10) A. Tuzhilin (*Stern School of Business, New York University, USA*), qui participent au comité de lecture de l'ouvrage QMDM, F. Guillet and H. Hamilton (eds), Springer 2006.

Partie IV : Enseignement

1. RESPONSABILITES PEDAGOGIQUES

1.1. Responsable des Relations Internationales

Au niveau du Département Informatique

Depuis septembre 2000, je suis *responsable des relations internationales* du département informatique de Polytech'Nantes. Ma fonction consiste à gérer et à développer l'ensemble des relations avec les universités et les entreprises étrangères en informatiques. Elle concerne 7 activités principales :

- l'établissement de nouvelles *conventions inter-universitaires* (Cantho-Vietnam [2003], Laval-Canada [2004], Wayne-USA [2005], IAI-Libreville-Gabon [en cours], Université du Piémond Oriental-Italie [en cours]) ;
- le maintien du *réseau de contacts* comportant 28 universités via des conventions inter-universitaires ;
- la gestion des offres de stage émanant d'entreprises étrangères (30-40 / an)
- la gestion des *semestres étudiants* vers l'étranger (5-10/an en 3ème année), et en provenance de l'étranger (2-5/an), à travers les programmes ERASMUS, CREPUQ, et EGIDE ;
- la gestion des *stages ingénieur* effectués à l'étranger en : 1ère année (5-10/an), 2ème année (20-25/an), et 3ème année (10-15/an) ; et en provenance de l'étranger (3-5/an) ;
- le maintien d'un *site intranet* de diffusion d'information ;
- le *recrutement* des étudiants en provenance de l'étranger.

Au niveau de la formation doctorale

Depuis 2002, je suis *chargé de la coordination internationale* pour le Master "Extraction de Connaissances dans les Données" (Master ECD). Depuis 2006, mon activité de coordination s'est étendue à la création du Master européens MDM&KD. (Voir 0, page 40).

Au niveau de l'Equipe COD

- la gestion des *stages et des semestres de Masters*, dans le cadre du Master ECD : étudiants de Bucarest (2-3/an), et de Cantho (2-3/an), à travers des programmes SOCRATES, EGIDE, et AUF

- la gestion des enseignants chercheurs invités à Polytech'Nantes (7 entre 2002 et 2006). Voir Partie III :8.6.1, page 59.

Types	***Années : sortants+entrants***						
	00/01	***01/02***	***02/03***	***03/04***	***04/05***	***05/06***	***06/07***
Stages 1A/2A	26+0	8+0	20+0	15+0	18+0	20+0	
Stages 3A	15+0	10+0	5+0	13+0	10+0	18+1	
Semestres	1+0	1+0	1+0	1+0	2+4	10+4	8+2
Stages Master		2+0	2+3	3+2	3+2	2+2	
Semestres Master				1		2	3
Conventions				1		2	2
Invités			2	2	2	3	
Total	**42+0**	**21+0**	**30+3**	**32+6**	**33+8**	**50+13**	**NC**

Récapitulatif des flux entrants/sortants à l'international, département informatique

1.2. Participation à la création de nouvelles formations

J'ai activement contribué à la création du cursus ingénieur en Informatique décisionnelle (ID), du Master Extraction de Connaissances dans les Données (ECD), et de son extension européenne (MDM&KD).

1.2.1. Création du cursus Ingénieur ID

Sous la direction d'Henri Briand, je me suis impliqué dans la construction du cursus « Informatique Décisionnelle ». Il s'agit d'une option lourde, sur 3ans à temps plein, 2000 h, proposant un nouveau profil d'ingénieur en informatique en réponse aux évolutions du secteur informatique qui se dessinent sur les 10 ans à venir, et avec la préoccupation de renforcer les synergies entre les besoins des entreprises et les apports de la recherche. Ce cursus est associé à un comité de pilotage constitué pour moitié de responsables informatiques et d'enseignants chercheurs. L'option ID a nécessité la conception de nouveaux enseignements en systèmes d'information avancés, en aide à la décision, en gestion des connaissances,…

En particulier, je me suis impliqué dans la conception de 4 modules d'enseignement, que depuis je coordonne et dans lesquels j'interviens comme enseignant :

- *IL1 et IL2 – Ingénierie du Logiciel*, 1ère année ID, 200 h/an, 8 intervenants [2002-…]
- *SC2 –Gestion des connaissances,* 1ère année ID, 60 h/an, 5 intervenants [2003-…]
- *SC3 et SC4 – Principes des systèmes déductifs et Extraction de Connaissances dans les Données*, 2ère année ID, 70 h/an, 6 intervenants [2004-…]
- *SC5 – web sémantique*, 3ère année ID, 24 h/an, 3 intervenants [2005-…]

1.2.2. Création d'enseignements en Master

J'enseigne dans deux Masters : Systèmes d'aide à la décision (Master SAD) et Extraction de connaissances dans les données (Master ECD).

J'ai participé au lancement de ces deux masters, avec une plus forte implication dans la création du second Master (DEA puis Master ECD). Actuellement, je participe pleinement au lancement du master européen MDM&KD. (Voir 0, page 40).

Après en avoir conçu le contenu pédagogique, j'interviens comme enseignant et coordonne 3 cours de Master :

- *Knowledge Quality in Data Mining* , cours de tronc commun, Master *MDM&KD*, 30 h/an, 3 intervenants [2006-…]
- *Gestion et déploiement des connaissances,* cours de recherche, Master ECD, 20 h/an, 3 intervenants [2004-…]
- *Extraction des connaissances dans les données,* DEA Informatique, puis Master SAD, Nantes, 24 h/an, 3 intervenants, [2002-…]

J'avais précédemment coordonné les cours suivants :

- *Extraction et Gestion des connaissances,* cours de recherche, Master ECD, 20 h/an, 3 intervenants [2002-2004]
- *Qualité des connaissances*, cours de recherche, DEA ECD, 30h/an, 4 intervenants [2000-2002]

1.3. Responsabilités d'équipements pédagogiques

Dans le cadre de la commission « logiciel et matériel »' j'ai été amené à participer, en liaison avec le service informatique, à la stratégie d'équipement informatique du département : achat de logiciel et de matériel, installation de salles pédagogiques. En particulier, je suis responsable de trois projets d'équipements pédagogiques lourds, que j'ai initié puis porté (montage de dossier de financement, achat, installation, puis gestion et maintenance).

- **Equipement pédagogique « réalité virtuelle », 15 k€.** Il s'agit d'un ensemble informatique financé par la région « Pays de la Loire », qui comporte une paire de vidéo-projecteurs polarisés permettant la restitution stéréoscopique en relief (3D immersive) sur un écran panoramique (3m×3m) à travers des lunettes ad hoc, un serveur de calcul orienté graphique, un gant de données, et des logiciel de synthèse et d'intégration d'environnements 3D. Cette salle est destinée aux enseignements ID2 « IHM avancées » sur les interfaces informatiques du futurs. Récemment, elle a aussi stimulé des applications en recherche dans le cadre d'un projet régional « aide à la décision ».

- **Equipement pédagogique « visio-conférence », 10 k€**. Il s'agit d'un ensemble informatique multimedia dédié au téléenseignement multi-sites en en direct : mode présentiel par visio-conférence. Il est constitué de deux stations, d'un ensemble vidéo avec caméra et vidéo-projecteur, d'un ensemble audio avec micros HF et enceintes 5.1, d'un tableau virtuel, et de logiciels multimedia. Cette salle, financée sur fonds propres, constitue un support majeur pour le téléenseignement en formation doctorale (Master ECD et MDM&KD) sur quatre sites en simultané (voir 0, page 40).

- **Equipement pédagogique « salle Informatique décisionnelle », 50k€**, financé par la région. Il s'agit de l'équipement d'une salle de travail complète pour les projets étudiants en cursus « Informatique décisionnelle » : 15 stations de travail, un serveur de calcul quadri-processeurs, un serveur de stockage, et une baie de stockage de 10 To.

1.4. Responsabilités de modules d'enseignements

Je suis responsable de plusieurs modules d'enseignement dans lesquels j'interviens comme enseignant principal (sauf SC3). Il s'agit ici de concevoir les contenus pédagogiques des cours et les TD/TP associés, de coordonner les intervenants, et de gérer les évaluations, …

Formations Ingénieur, Département Informatique, Polytech'Nantes

- *IB1 et IB2 – Informatique de base*, 1ère année SILR, 200 h/an, 8 intervenants [1997-…]
- *IL1 et IL2 – Ingénierie du Logiciel*, 1ère année ID, 200 h/an, 8 intervenants [2002-…]
- *SC2, SC3, SC4, SC5 – Systèmes à base de* ID, 160 h/an, 8 intervenants [2003-…]
- *ECD5 - Extraction de connaissances dans les données*, 3ème année, 24 h/an, 3 intervenants [2001-2004]

Formation continue

- *Langages de programmation*, DUTIL (Bac+4), 60h/an, 3 intervenants [1997-2002]

Formations extérieures

- *Extraction de connaissances dans les données*, IUP GIS- Vannes (Bac+5), 30h/an, 3 intervenants [1999]

Autres formation Ingénieur

- *Traitement et transmission de l'information*, Ecole Royale de l'air, Marrakech (Bac+3 et bac+5), 150h/an, 3 intervenants. Responsable de la formation des élèves officiers (pilotes et transmissions) en traitement et transmission de l'information pendant 2 promotions [1999-1991]

1.5. Autres responsabilités dans le département Informatique

- Membre nommé du *conseil pédagogique* [2004-…]. Participation à la cohérence pédagogique inter-modules, des enseignements, …
- Membre de *la commission "matériels et logiciels" [2000-…] :*

2. ENSEIGNEMENTS DISPENSES

2.1. Production de supports pédagogiques

La nature des supports pédagogiques varie selon la nature des cours dispensés. Il s'agit soit de documents rédigés, soit de recueils de transparents très détaillés, soit plus rarement de notes de recherche. En particulier tous les TD et TP ainsi que les cours à caractère technique (algorithmique et langage C, Unix, Java, Smalltalk) sont associés à des documents complets et rédigés. Je ne dispense pas de cours sans distribuer au préalable un support pédagogique papier aux étudiants.

2.2. Troisième Cycle

Mes interventions ont eu lieu dans Masters :

- Le DEA puis Master ECD cohabilité par Lyon 2, Paris 11 et Polytech'Nantes

- Le Master européen « Master in Data Mining and Knowledge Discovery » (MDM&KD), extension internationale du master ECD.

- Le DEA informatique puis Master SAD de l'université de Nantes

En synthèse, ces enseignement peuvent être classées en deux catégories : (1) *Extraction de Connaissances dans les Données* (ECD), et (2) *Gestion des Connaissances*.

Module	**Niveau**	**Lieu**	**Type**	**Années**	**H/an**
Qualité des connaissances	DEA ECD	Polytech'Nantes	Cours	00-02	5
Extraction et Gestion des Connaissances	DEA ECD	Polytech'Nantes	Cours	02-04	6
Gestion et déploiement des Connaissances	Master ECD	Polytech'Nantes	Cours	04-…	12
Extraction de Connaissances (ECD)	DEA Info	Univ. Nantes	Cours	02-04	6
ECD – fouille de données, qualité	Master SAD	Univ. Nantes	Cours	04-…	9
Knowledge Quality in Data Mining	Master MDM&KD	Univ. Nantes	Cours	06-…	15

2.3. Formation Ingénieur

Je suis intervenu dans trois formations ingénieur différentes :

- de 1989 à 1991 à *l'Ecole Royale de l'Air* de Marrakech (ERA, Bac+5), école d'officiers de l'armée de l'air affiliée à l'école de l'air de Salon de Provence ;

- de 1991 à 1995 à *l'Ecole nationale Supérieure des Télécommunications de Bretagne* (ENST Bretagne) ;

- et enfin depuis 1995 à *l'Ecole polytechnique de l'université de Nantes* (Polytech'Nantes).

J'ai aussi assuré un enseignement dans le *mastère Image et Intelligence Artificielle* de l'ENST Bretagne.

L'ensemble des cours dispensés peut être synthétisé en 4 domaines :

- *Informatique système et algorithmique* : architecture des systèmes, systèmes d'exploitation, Parallèlisme et clusters, langages et compilateurs, C et Unix, algorithmique

- *Traitement et transmission de l'information* : théorie du signal, transmission du signal, réseaux et télécommuniactions, sécurité

- *Génie Logiciel* : Programmation objet, IHM, Smalltalk et Java

- *Systèmes d'information* : fouille de données, gestion des connaissances, web sémantique.

3ème année (bac+5) et mastère (bac+6)

Module	Niveau	Lieu	Type	Années	H/an
Smalltalk	Mastère IIA, Bac+6	ENST Bretagne	Cours/TP	92-95	20/20
Mesure de qualité en ECD	Ingénieur, Bac+5	ENST Bretagne	cours	00-02	3
Fouille de données - ECD	Ingénieur, Bac+5	Polytech'Nantes	Cours/TP	01-05	18/6
Parallélisme - Clusters	Ingénieur, Bac+5	Polytech'Nantes	Cours/TP	01	9/9
Réseaux - Sécurité	Ingénieur, Bac+5	Polytech'Nantes	Cours/TP	01	9/9
Web Sémantique	Ingénieur, Bac+5	Polytech'Nantes	Cours/TP	05-…	3/6

2ème année (bac+4)

Module	Niveau	Lieu	Type	Années	H/an
Théorie du signal	Ingénieur, Bac+4	ERA	Cours	89-91	50
Smalltalk	Ingénieur, Bac+4	ENST Br	TP	92-95	20
Langages et compilateurs	Ingénieur, Bac+4	ENST Br	TP	92-95	20
Extraction de Connaissances (ECD)	Ingénieur, Bac+4	Polytech'Nantes	Cours/TP	04-…	12/6

1ère année (bac+3)

Module	Niveau	Lieu	Type	Années	H/an
Théorie du signal	Ingénieur, Bac+3	ERA	Cours/TP	89-91	50/50
Smalltalk	Ingénieur, Bac+3	Polytech'Nantes	TP	95-97	12
Unix et Syst. Exploitation	Ingénieur, Bac+3	Polytech'Nantes	TP	95-97	50
Architecture des Systèmes	Ingénieur, Bac+3	Polytech'Nantes	Cours/TP	97-…	40/50
Programmation objet : java, C++	Ingénieur, Bac+3	Polytech'Nantes	TP	97-…	40
IHM	Ingénieur, Bac+3	Polytech'Nantes	Cours/TP	97-01	8/12
Traitement de l'information	Ingénieur, Bac+3	Polytech'Nantes	TD	97-00	12
Réseaux et Télécommunications	Ingénieur, Bac+3	Polytech'Nantes	TD	00-02	12
Gestion des Connaissances	Ingénieur, Bac+3	Polytech'Nantes	Cours/TP	03-…	12/6

2.4. Formation Continue

Je suis intervenu 1995 à 2002 dans la formation continue diplômante DUTIL (Formation chef de projet en informatique à Bac+4) de Polytech'Nantes, et en stages de formation pour les entreprises. Les cours dispensés ont concerné les fondamentaux de l'informatique : *systèmes*, *algorithmique*, *programmation objet*.

Module	Niveau	Lieu	Type	Années	H/an
Langage C	Stage FC, Bac+4	Polytech'Nantes	Cours/TP	95-97	20/15
Unix	Stage FC, Bac+4	Polytech'Nantes	Cours/TP	95-97	20/10
Unix	Dutil, Bac+4	Polytech'Nantes	Cours/TP	95-97	10/14
Algorithmique et langage C	Dutil, Bac+4	Polytech'Nantes	Cours/TP	95-97	10/14
Smalltalk	Dutil, Bac+4	Polytech'Nantes	Cours/TP	95-02	10/10
Algorithmique et langage C	Dutil, Bac+4	Polytech'Nantes	Cours/TP	95-02	10/20
Java	Dutil, Bac+4	Polytech'Nantes	Cours/TP	97-02	6/9

3. ENCADREMENT DE PROJETS ET DE STAGES

Une part importante de mon activité pédagogique est consacrée à l'encadrement de projet et de stages de fin d'étude dans trois cursus : le cycle C Ingénieur en informatique du Cnam, le cycle ingénier du département informatique de Polytech'Nantes, et la formation continue diplômante DUTIL.

3.1. Ingénieurs du Conservatoire des Art et Métiers (Cnam)

3.1.1. Mémoires Cnam

Ces stages de fin d'études, d'une durée de 9 à 12 mois, achèvent et valident le diplôme « Ingénieur en informatique », Cycle C, du Conservatoire des Art et Métiers de Nantes.

1. P. Fargeaud (1997). Reconnaissance d'image par analyse spectrale des contours.
2. L. Jaffrelot (2004). Développement d'un atelier pour la validation des règles extraites dans des bases de données.
3. M. Ahlonsou (2005). Transformation des concepts des diagrammes de uml en full owl.
4. N. Ronarc'h (2005). La Modélisation des Connaissances Emotionnelles d'Agents Autonomes.
5. R. Dailly (2005). Web sémantique et e-learning - Ontologie pédagogique d'une plate-forme d'apprentissage.
6. Y. Fosset (2006). Interactions entre agents émotionnels.
7. X. Aimé (2006). Représentations graphiques pour l'alignement interactif d'ontologies.

3.1.2. Probatoires Cnam

Il s'agit de projets d'études bibliographiques, d'une durée de 5 semaines, validant le passage en dernière année du Cycle C « Ingénieur en informatique » du Conservatoire des Art et Métiers de Nantes.

1. N. Neilz et D. Vitige (2000).Amélioration de l'interface du logiciel Samanta. Projet Cnam, 4 mois.
2. P. Moreau (2004).RDF : un outil de l'informatique décisionnelle pour la gestion des connaissances - enjeux et perspectives. Probatoire Cnam.
3. M. Ricaud (2005). Les Graphes conceptuels pour les ontologies. Probatoire Cnam.
4. F. Rouault (2005). Ontologies et WEB sémantique. Probatoire Cnam.

3.2. Masters

Voir Partie III :4.2 DEA et MASTER, page 46.

3.3. Ingénieurs Polytech'Nantes

Dans le cadre du département Informatique de Polytech'Nantes, j'encadre de nombreux projets depuis 1997.

Projets de 2ème année (Bac+4)

Il s'agit de projet d'une durée 300 heures (1 journée par semaine sur toute l'année) menés en partenariat avec une entreprise. J'encadre **3 à 4 projets de 2ème année par an** depuis 1997.

Projets de 3ème année (Bac+5)

Il s'agit de projet d'une durée 150 heures (1 journée par semaine sur un semestre) menés dans un contexte soit d'entreprise, soit de recherche. J'encadre **2 à 3 projets de 3ème année par an** depuis 1997.

Stages de fin d'étude (Bac+5)

Le dernier semestre d'étude est réalisé en stage dans une entreprise (5 mois). J'ai directement encadré **5 stages d'ingénieur** [99-04], dans le cadre des contrats Samanta et Athanor avec la société PerformanSe SA, Carquefou.

D'autre part, un tuteur enseignant est affecté au suivi de chaque stage (encadrement et visite). Je suis le tuteur de **4 à 5 Stages de fin d'étude par an** depuis 1997.

3.4. Formation continue DUTIL

La formation continue DUTIL « Chef de projet en informatique » de Polytech'Nantes est une formation présentielle diplômantes (Bac+4) d'une durée d'un an, incluant un projet de 300 heures. J'ai encadré **1 à 2 projets DUTIL par an** entre 1997 et 2002.

3.5. Autres projets Ingénieurs

1. Ingénieurs de l'Ecole Nationale Supérieure des Télécommunications de Bretagne

Entre 1992 et 1995, j'ai encadrés **3 projets d'étudiant** au département informatique (IASC) de l'Ecole Nationale Supérieure des Télécommunications de Bretagne, en 1ère, 2ème et 3ème année (Bac+3 à Bac+5).

2. Officiers ingénieurs de l'Ecole Royale de L'air

Entre 1989 et 1991, j'ai encadrés **4 projets d'élève officier** à l'Ecole Royale de l'Air (ERA, 2 mois, 150 heures, Bac+5), en électronique et en informatique

Types	***Années***										
	…/ 97	***97/ 98***	***98/ 99***	***99/ 00***	***00/ 01***	***01/ 02***	***02/ 03***	***03/ 04***	***04/ 05***	***05/ 06***	***06/ 07***
Projets 2A (300h)	3	3	2	3	3	5	5	4	4	2	2
Projets 3A (150h)	2	2	2	1	3	5	2	2		3	2
Tutorat 3A (avec visite)	6	6	6	5	5	5	5	4	5	5	
Stages 3A (5 mois)	1	1	1	1	1	2	2	2	3		
Stages DUTIL (300h)	1	1	1	1	2	1					
Stages CNAM (9 mois)		1		1				1	3	3	2
Stages Master (4 mois)			1	3	1	2	3	3	4	1	
Soutenances thèses					2					1	1
Autres	7				2					1	1
Total	**20**	**14**	**12**	**15**	**17**	**20**	**17**	**16**	**19**	**15**	

Récapitulatif des encadrements d'étudiants

Résumé : Les travaux synthétisés dans ce document s'insèrent dans deux domaines de recherche : l'Extraction de Connaissances dans les Données (ECD), et la Gestion et l'Ingénierie des Connaissances (GIC). Ils sont fédérés autour d'un objectif commun, celui du traitement des connaissances, et partagent une préoccupation sous-jacente pour l'aide à la décision et pour le développement d'applications.
Nos travaux en ECD ont porté sur la phase terminale du processus de fouille de données, celle du *post-traitement* des connaissances par un décideur/expert des données. Plus particulièrement, nous nous sommes intéressés au modèle des *règles d'association* qui permet la découverte non supervisée de tendances implicatives au sein des données. Les avantages de ce modèle sont contrebalancés par un inconvénient majeur, celui de délivrer une quantité prohibitive de règles. Afin de contourner cette limite au post-traitement, nous proposons coupler trois approches envisagées séparément dans la littérature : les mesures de qualité, les représentations graphiques, et l'interactivité.
Nous avons proposé un recensement et un classement des *mesures de qualité* disponibles, conçu de nouvelles mesures, développé une approche contextuelle pour l'analyse corrélative des mesures par des graphes de corrélation. Cette panoplie d'outils et de méthodes offre un ensemble de solutions afin de guider l'utilisateur vers le choix des meilleures mesures, et in fine des meilleures règles.
Afin d'améliorer le post-traitement, nous avons choisi de coupler les mesures de qualité à des représentations graphiques interactives adaptées au traitement de grands volumes de règles. Nous avons proposé une *approche anthropocentrée*, fondée sur une stratégie cognitive de *ciblage de règles*, qui permet à l'utilisateur de mener une véritable *fouille de règles* dans la masse des règles produites. Cette approche, qui concilie intelligibilité et efficacité grâce à des algorithmes locaux d'extraction, a été déclinée en deux représentations complémentaires, la première par des *graphes dynamiques de règles*, et la seconde par des *métaphores 3D en Réalité Virtuelle*.
Nos travaux en GIC ont concerné le traitement des connaissances par des approches issues du *web sémantique*. Nous avons conçu une approche *serveur de connaissances* pour les mémoires organisationnelles, une modélisation des connaissances pour des *agents émotionnels*, et une technique originale d'*alignement d'ontologies*.
L'approche *serveur de connaissances* propose une analogie avec un serveur web transposant ses services à la connaissance. Cette approche, implémentée dans l'outil SAMANTA destiné à la gestion des diagnostics de maintenance, est fondée sur trois ontologiques complémentaires. L'outil incorpore un éditeur graphique d'ontologies, des ressources multimédia (dont des modèles en réalité virtuelle), et enfin un raisonneur prolog. Nous avons ensuite étendu et généralisé cette approche dans l'outil Athanor.
Afin d'intégrer la dimension émotionnelle dans un système d'aide à la décision sur la dynamique des groupes, nous avons proposé une modélisation agent UML et proposé un modèle EST pour des *agents émotionnels* BDI. Cette modélisation permet de produire une base de connaissances en RDF/OWL décrivant l'évolution interne des agents, et leurs interactions.
Enfin, en nous inspirant de nos travaux sur les règles d'association, nous avons proposé une méthode d'*alignement d'ontologies*. Plus précisément, nous considérons qu'une ontologie est un graphe orienté de concepts structurés par une relation de subsomption, où chaque concept est décrit dans des documents textuels. Notre approche se décompose alors en deux étapes consécutives : l'extraction terminologique ; puis l'extraction d'un ensemble minimal, selon un critère de réduction des redondances, de règles d'association entre concepts.

Mots-clés : *Fouille de données, Mesures d'intérêt, Visualisation de règles, Fouille de règles, Réalité virtuelle, Gestion des connaissances, Serveur de connaissances, Agent émotionnels, Alignement d'ontologies, Aide à la décision.*