Fouille de textes : des méthodes symboliques pour la construction d'ontologies et l'annotation sémantique guidée par les connaissances

Yannick Toussaint

Nancy-Université
Université Henri Poincaré

Université Henry Poincaré, Nancy 1
Département de formation doctorale en informatique **École doctorale IAEM Lorraine**
UFR STMIA

Fouille de textes : des méthodes symboliques pour la construction d'ontologies et l'annotation sémantique guidée par les connaissances

Projet de Recherche

présenté et soutenu publiquement le 21 novembre 2011
en vue de l'obtention d'une

Habilitation à diriger les recherches de l'Université Henri Poincaré – Nancy 1

(spécialité informatique)

par

Yannick Toussaint

Composition du jury

Rapporteurs : Daniel Kayser, Professeur, Université Paris-Nord 13, LIPN
Stan Matwin, Professeur, Université d'Ottawa, Canada
Pierre Zweigenbaum, DR CNRS, LIMSI

Examinateurs : Nathalie Aussenac, DR CNRS, IRIT
Béatrice Daille, Professeur, Université de Nantes, LINA
Amedeo Napoli, DR CNRS, LORIA
Jean-Marie Pierrel, Professeur, Université UHP-Nancy 1, ATILF

Loria
Laboratoire lorrain de recherche en informatique et ses applications

Inria informatiques mathématiques

Mis en page avec la classe thloria.

Table des matières

Projet de recherche

Chapitre 1

La fouille de textes

Ce mémoire présente mon projet de recherche sur la fouille de textes reposant sur des méthodes symboliques de fouille de données. Pourquoi des méthodes symboliques ? Elles permettent, par exemple, des raisonnements reposant sur la logique, des explications ou des définitions plus facilement interprétables par des experts, et permettent aussi d'envisager un enrichissement successif des résultats au cours des différentes itérations. La fouille de textes, traduit du terme anglais *text mining*, est apparue dans la deuxième moitié des années 90, en écho aux travaux réalisés depuis les années 80 sur des bases de données. Cependant, selon la culture scientifique dont sont issus les chercheurs, le terme de fouille de textes désigne différents types d'activités. Je considère principalement la fouille de textes comme un paradigme de l'extraction de connaissances à partir de données pour lequel les données sont des textes.

Mes recherches visent donc à proposer des méthodes et outils pour permettre de raisonner sur de grands volumes de textes. Pour des chercheurs issus de la **Recherche d'Information (RI)**, raisonner, ce serait collectionner de très grandes quantités de documents et être capable de les manipuler, de les classer, ou de les retrouver à la demande : chercher des cooccurrences et les filtrer pour ne garder que certaines d'entre elles. Pour d'autres, raisonner c'est représenter un énoncé en langue naturelle par une formule logique et exploiter cette formule avec les outils de preuve que cette logique propose. Les travaux en **Traitement Automatique de la Langue (TAL)** s'inscrivent dans cette perspective. Du côté de la RI, les outils sont polyvalents, clés en main et très robustes. Mais dès que l'on envisage une certaine forme de compréhension de la langue naturelle, l'outil doit être taillé sur mesure, en fonction d'une analyse des besoins approfondie. Enfin, le domaine de **l'Extraction de Connaissances à partir de Bases de Données (ECBD)** vise, quant à lui, à construire une conceptualisation du domaine à partir des données.

Ce chapitre esquisse un positionnement de ces trois domaines par le biais d'un historique très rapide avec une volonté non dissimulée de donner une vision subjective de ces domaines. L'objectif est de montrer que la fouille de textes a été abordée par des domaines différents, avec des objectifs différents. Les contributions de ces domaines sont cependant cohérentes au regard de ce que devrait être un processus "complet" de fouille de textes et c'est un point sur lequel je reviens dans la conclusion de ce mémoire. Il est intéressant d'observer que la fouille de textes pose à ces différents domaines des défis qui les obligent progressivement à coopérer, permettant par exemple d'emprunter à la recherche d'infor-

mation la robustesse de ses processus et au traitement de la langue, l'identification de relations sémantiques... Dans les chapitres qui suivent cette introduction, je reprendrai de façon plus argumentée les points importants pour mon projet de recherche.

1.1 Positionnement

Note de lecture : Pour garantir l'indépendance de la lecture de chacun des deux documents soumis pour l'HdR, cette section 1.1 est reprise sous forme d'un chapitre dans le document détaillant mon bilan scientifique. S'il l'a déjà lu, je suggère au lecteur de passer à la section 1.2 de ce document.

1.1.1 Données, informations et connaissances

Avant toute chose, il est important de préciser quelques définitions car il y a souvent confusion entre ce que sont les données, les informations et les connaissances. J'adopte donc les définitions proposées par Schreiber *et al.*[SAA+99] :

Une **donnée** est un signal brut, non interprété. Une donnée peut être issue d'un capteur, par exemple, un thermomètre ou issue d'environnement comme le web ou des bases de données. Elle peut être faiblement structurée comme peuvent l'être certains textes ou plus fortement structurée comme peuvent l'être les documents codés en XML.

Une **information** est une donnée qui a été interprétée. Par exemple, le dépassement d'un seuil de température est une information. Un texte rédigé dans une langue étrangère que je ne maîtrise pas est une donnée et n'est que suite de mots ou de signes. En revanche, je vais interpréter un texte écrit en français et il devient donc une source d'information.

Une **connaissance** est un ensemble de données et d'informations mises en œuvre pour assumer une tâche ou produire une nouvelle information. La connaissance est souvent associée à deux aspects : la *finalité* car la connaissance est mise en œuvre pour atteindre un objectif et sa *capacité générative* puisque la connaissance permet de produire de nouvelles informations.

1.1.2 La recherche d'information

La recherche d'information (RI) se définit par un ensemble de méthodes et d'outils qui permettent à un utilisateur de formuler une requête (*i.e.* un ensemble de critères) et qui sélectionnent dans un fond documentaire les documents répondant à ces critères. Les travaux en recherche d'information automatisée remontent aux années 50 et certains ouvrages [vR79, Sal89], un peu anciens, font référence dans le domaine. Cependant, comme le montre [MRS08], les enjeux exposés dans ces ouvrages restent toujours d'actualité : l'analyse ou la caractérisation automatique de documents, la classification automatique (supervisée ou non supervisée), la structure des documents et des fichiers, les stratégies de recherche et l'évaluation des résultats. Le domaine de recherche est très actif et des avancées importantes sont à souligner. Ainsi, le codage de l'information contenue dans un document peut se faire au niveau des mots contenus dans le document, des termes d'un domaine, et peut prendre en compte la structure, par exemple XML, du document.

Les résultats d'une requête peuvent se présenter sous forme d'une liste ou d'une liste structurée (classes, hiérarchie...). Il existe en effet de nombreuses méthodes de classification comme les *k-means* ou les cartes auto-organisatrices de Kohonen visualisant les classes de documents sur des cartes. Les méthodes de classification et de cartographie appliquées aux mots-clés ou aux termes d'un domaine sont assez souvent qualifiées d'outil de fouille de textes comme le montre, entre autres, l'ouvrage édité à partir du workshop "Text Mining" associé à la conférence *SIAM International Conference on Data Mining (SDM 2002)* [Ber04]. De ce fait, il y a souvent confusion entre d'une part, la classification (non supervisée) de textes, la catégorisation (classification supervisée) de textes et l'identification de thèmes, ou encore la détection de tendances avec, d'autre part, le processus de construction de connaissances à partir de textes dont ils peuvent éventuellement faire partie.

Dans un système de RI, un point important est la capacité à discriminer les documents. L'objectif premier de la recherche d'information est de pouvoir proposer en réponse à une requête les documents les plus pertinents. L'enjeu majeur ne se situe donc pas dans la « compréhension » d'un document mais sur la capacité du système à distinguer un document d'un autre. La présence d'un mot, d'un mot-clé ou d'un terme dans un document peut ainsi être pondérée, par exemple, par le TF-IDF (Term Frequency - Inverse Document Frequency) reflétant son pouvoir discriminant. Dans une problématique de compréhension, la pondération peut avoir des effets trop réducteurs. Barthes [Bar66] soulignait que « Structuralement, le sens ne naît point par répétition mais par différence, en sorte qu'un mot rare, dès lors qu'il est saisi dans un système d'exclusions et de relations, signifie tout autant qu'un terme fréquent. » Des pondérations du type TF-IDF sont pourtant utilisées par certains extracteurs de termes pour que les résultats de ces systèmes "très bavards" aient une taille analysable par un humain.

Le second point important est la robustesse face au volume de documents. Le volume de documents, qu'il s'agisse de bases documentaires, de bases documentaires sur le web, ou du web en général, ne fait que croître. Les méthodes utilisées sont essentiellement statistiques ou probabilistes et doivent pouvoir s'appliquer à de gros volumes de données. Les temps de réponse doivent être courts et les méthodes de représentation vectorielle des documents sont, de ce point de vue, performantes.

Les méthodes utilisées en recherche d'information sont peu sensibles au domaine sur lequel on les applique, d'où une contextualisation faible. La contextualisation existe cependant pour certains systèmes par la prise en compte de connaissances du domaine ou au travers de la construction de réseaux (notamment des réseaux sociaux). Cependant, cette contextualisation peut faire appel à des algorithmes d'apprentissage et reste très superficielle en comparaison avec celle qui est nécessaire pour le traitement automatique de la langue ou pour l'extraction d'information.

Enfin, une "tradition" forte en RI est l'évaluation des résultats. L'évaluation de la qualité d'un système de RI se fait essentiellement par trois mesures : le rappel, la précision et une combinaison de ces deux premières, la f-mesure. Un système de RI idéal propose en réponse à une requête tous les documents pertinents présents dans la base (rappel élevé) et très peu de documents non pertinents (précision élevée).

1.1.3 Le traitement automatique de la langue

Le Traitement Automatique de la Langue (TAL) est un domaine incontournable dès lors qu'il s'agit d'extraire des éléments de sens à partir de textes. Le TAL [JM00] se positionne à la frontière entre la linguistique et l'informatique. Les applications de ces travaux sont assez diverses : la traduction automatique (parmi les premiers travaux dans les années 50), la génération automatique de textes, la représentation du contenu d'une phrase ou d'un discours ... C'est ce dernier point qui nous concerne plus particulièrement : comment extraire et représenter le sens d'un énoncé. La plupart de ces travaux s'intéressent à modéliser, parfois de façon très fine (les quantifieurs généralisés...) les phénomènes de langues dans un cadre logique et propose de représenter la sémantique d'une phrase ou d'un discours sous une forme logique (en lambda calcul) comme par exemple dans la DRT (Discourse Representation Theory)[Kam81].

Dans l'idéal, on aurait aimé que la contextualisation – le rôle du domaine – pour de telles applications soit faible et pouvoir ainsi représenter le sens d'un énoncé dans l'absolu : il s'agirait en quelque sorte d'une traduction d'une forme linguistique en une forme logique. La réalité est cependant bien différente en raison de la complexité de la langue (ambiguïtés, anaphores, structures syntaxiques complexes...), des types de textes (lettre, article scientifique, manuel...) et de la diversité des domaines (médecine, astronomie...).

Les méthodes statistiques peuvent être couplées à des méthodes d'apprentissage et peuvent parfois être gage de robustesse dans le traitement. Ces méthodes statistiques sont fortement présentes dans des applications du traitement de la langue comme les traitements de la parole ou la traduction automatique, mais l'analyse sémantique de phrases ou de discours repose très majoritairement sur des processus symboliques faisant appel à des grammaires et opérateurs logiques utilisés pour la construction compositionnelle du sens. Cependant, les étiqueteurs morphosyntaxiques exploitant des approches statistiques dans les années 90 font émerger l'idée d'analyseurs syntaxiques robustes et partiels et ont relancé bon nombre de travaux en linguistique de corpus. Ces étiqueteurs constituent aujourd'hui la couche basse de toute chaîne de traitement. Des mesures statistiques peuvent également être associées aux règles de grammaire pour réduire le nombre d'arbres syntaxiques produits et de gros progrès en terme de couverture et de robustesse ont été réalisés sur l'étape d'analyse syntaxique. Les modèles de langage ont aussi été utilisés pour le typage sémantique de mots (en traitement de la parole) mais cela ne permet pas leur intégration dans un système de raisonnement.

Des environnements complets d'analyse syntaxico-sémantique sont progressivement développés : le programme anglais ANLT (Alvey Natural language Tools) dans la fin des années 80, le NLTK (Natural Language ToolKit), les outils proposés par le « Stanford NLP Group», ou encore le C&C qui intègre également la construction des DRS (Discourse Representation Structures) et un démonstrateur permettant de raisonner sur les textes. Mais, à supposer donc que l'analyse syntaxico-sémantique d'un énoncé soit résolue, il subsiste deux difficultés majeures lorsque l'on s'intéresse à la représentation des connaissances. La première est liée aux connaissances implicites nécessaires à la compréhension d'un texte. La seconde est liée au fait que l'analyse d'un énoncé « colle » à la forme linguistique. Si transformer une phrase passive en une phrase active est assez "simple", il est des phénomènes de paraphrase beaucoup plus difficiles à résoudre. Par exemple, un

cuisinier expert aimerait que les deux phrases suivantes aient une représentation sémantique proche puisque le résultat de l'exécution de ces deux instructions conduit à deux situations très proches : *« Heat ghee in a large soup pot.»* et *« Melt butter in a large stock pot over a medium heat.»*

Enfin, sans doute en raison de la diversité des objectifs conduisant à l'utilisation des systèmes de TAL, l'évaluation reste une pratique encore assez peu courante dans le TAL comparée à la RI, mise à part sur les couches basses comme l'étiquetage morphosyntaxique (projet GRACE) et quelques tentatives au niveau terminologique. Au niveau sémantique, le TAL recourt à une évaluation qualitative visant à identifier les phénomènes de langue correctement traités et ceux qui génèrent des erreurs, par exemple, dans la forme logique.

Rencontre entre TAL et RI : l'extraction d'information

La Recherche d'Information et le Traitement Automatique de La Langue sont deux domaines de recherche qui ont évolués séparément et même s'ils ont tous deux un ancrage dans l'informatique, l'un était davantage tourné vers le documentaire et l'autre vers la linguistique. La première rencontre entre ces deux domaines, bien qu'encore timide, s'est produite dans le cadre des travaux sur l'Extraction d'Information (EI) [Poi03]. L'EI est née d'une initiative de la DARPA aux États-Unis en 1987. Elle consiste à rechercher dans des textes en langue naturelle des informations à propos de classes d'entités et de relations prédéfinies et de stocker cette information dans un modèle ou dans une base de données. L'idée initiale était de pouvoir repérer dans les dépêches journalistiques les situations d'attentat, les acteurs (victimes, agresseurs) et les modes d'action. Elle a été ensuite appliquée à des domaines très divers. L'EI peut se voir comme le fait de compléter un formulaire – une sorte de *frame* prédéfini – ou de remplir une base d'informations structurée à partir de texte libre.

Le développement d'un système d'extraction d'information repose sur trois grandes étapes [NVB01] : (i) il faut identifier les fragments de textes pertinents, *i.e.* contenant une information ; (ii) définir la structure de représentation de l'information, puis (iii) développer les règles permettant d'identifier l'information puis remplir la structure proposée. Le principe sous-jacent à l'extraction d'information est de décomposer l'identification d'informations, parfois complexes, en des sous-problèmes simples. Ainsi, les grandes tâches qui sont classiquement identifiées sont [HDG00] :

- Reconnaissance d'entités nommées : noms de personnes, d'organisations, des expressions temporelles comme les dates ou les heures, unités de mesure...
- Résolution des coréférences : pronoms, la désignation par la fonction (Premier Ministre), abréviations et variantes orthographiques (béta-lactamase, $\beta-$lactamase)...
- Extraction de propriétés : localisation géographique, association entre une personne et ses fonctions...
- Identification de relations : interaction entre des protéines, personne employée par une société...
- Identification des événements : les événements sont des relations complexes telle que la description d'une attaque terroriste dans laquelle doivent être identifiés les terroristes, les victimes, les lieux, la date...

La robustesse des traitements est le point fort de cette approche. Les patrons sont

construits à partir d'éléments lexicaux, syntaxiques, des types sémantiques, mais également des indications de séquentialité dans le texte... Les patrons font généralement l'objet d'une évaluation de leur pertinence sur des corpus de tests. L'approche par patrons est robuste et permet de traiter des phrases longues et syntaxiquement complexes. C'est donc une approche incontournable dans le cas de phrases complexes pour lesquelles les risques d'erreur par des approches à base de grammaire sont trop élevés.

Développer à la main un ensemble de règles pour l'extraction d'information a un côté ludique. La convergence (si convergence il y a) vers un ensemble de règles stable et exhaustif par rapport à un objectif donné est lente et après deux années de recherche, il est parfois difficile de garantir la cohérence de l'ensemble des règles. De plus, les patrons définis sont très dépendants du domaine, des types de textes, et même de l'application visée. Définir un ensemble de patrons pertinents nécessite des compétences linguistiques, informatiques mais également une interaction suivie avec les experts du domaine des textes. Les premières tentatives pour automatiser l'apprentissage en EI remontent à 1993 avec les travaux de Riloff [Ril93] et, malgré les travaux en cours, notamment le réseau d'excellence européen PASCAL, la configuration d'un système d'EI est un travail d'ingénierie long et complexe.

La terminologie

Il serait un peu exagéré de relier les travaux de ces dix dernières années en terminologie à la recherche d'information. Il s'agit là d'une facette du TAL (mais pas uniquement du TAL) qui est essentielle pour nos travaux : approche sur corpus, empruntant à la linguistique de corpus des méthodes partielles et robustes, exploitant statistiques et recherche de patrons syntaxiques.

La terminologie joue un rôle essentiel dans l'accès au contenu des textes dans des domaines spécialisés. Le terme peut être vu comme une trace linguistique du concept mais, dans les travaux récents en terminologie, le terme, tout comme le concept, ne préexiste pas, il est construit à partir des manifestations linguistiques en corpus qui, comme le souligne N. Aussenac [AG05], permettent de définir le concept.

La terminologie peut être vue comme un nouveau point de rencontre entre la RI, le TAL mais également avec la représentation des connaissances. Plus fortement ancrée du côté du TAL, l'identification des termes fait grandement appel à une caractérisation linguistique de ce qu'est un terme. Ainsi des patrons syntaxiques – <*Nom*> de <*Nom*>, par exemple – sont proposés comme étant potentiellement des structures de termes, appelés candidats-termes. En revanche, ces patrons sont bruités, au sens où les candidats termes ainsi repérés ne correspondent pas tous à des termes. Les patrons syntaxiques sont généralement associés à des mesures statistiques d'occurrence dans des corpus de textes pour ne garder que ceux qui sont le plus probablement des termes.

Le terme est une entité linguistique et à ce titre, il subit des variations. Les outils comme FASTR ou ACABIT prennent en compte les variations des termes, variations linguistiquement motivées et qui peuvent ainsi augmenter le nombre de termes repérés dans les textes de 15 à 20%.

Dans nos travaux, nous utilisons divers outils de TAL pour identifier le contenu des textes, selon les besoins de l'application, le domaine ou les types de textes : des analyseurs

syntaxico-sémantiques, des outils d'extraction d'information ou des outils terminologiques.

Les domaines de la génomique ou de la médecine ont été et sont encore des cadres privilégiés d'application tant du TAL que de l'EI ou de la terminologie. Les phénomènes langagiers spécifiques à ces domaines ont été beaucoup étudiés et de nombreuses ressources sont disponibles rendant les outils plus performants. Certains travaux assimilent la fouille de textes à l'extraction d'information, tout au moins lorsqu'il s'agit d'apprendre les patrons d'extraction [MN03] ou de construire des classes d'objets linguistiques (verbes) en fonction de leur comportement dans les textes. Les experts d'un domaine ne sont pourtant pas intéressés, en premier lieu, par les connaissances linguistiques extraites de leurs textes, même si elles sont nécessaires pour les étapes ultérieures. Le TAL est souvent considéré dans un processus de fouille de textes comme l'étape permettant de passer du texte à des données pour que des processus de fouille de données puissent être appliqués. Nous verrons que ce schéma atteint vite ses limites et qu'une autre conception intégrant ces deux niveaux doit être proposée.

1.1.4 L'extraction de connaissances à partir de bases de données (ECBD)

En 1991, Piatetsky-Shapiro introduit comme titre de son ouvrage [PSF91] le terme de "Knowledge Discovery from Databases", abrégé par la suite en KDD et, dont l'équivalent français est Extraction de Connaissances à partir de Bases de Données (ECDB). Ce n'est que vers 1995 que l'usage des termes Knowledge Discovery from Databases et Data Mining se précise. L'extraction de connaissances à partir de bases de données désigne alors le processus global de découverte de connaissances qui permet de passer de données brutes à des connaissances alors que la fouille de données n'est qu'une étape de l'ECBD au cours de laquelle un modèle est construit.

La définition de l'ECBD a été enrichie au cours du temps et celle proposée par Fayyad[FPSS96] est maintenant consensuelle : *L'extraction de connaissances à partir de bases de données est un processus non trivial qui construit un modèle valide, nouveau, potentiellement utile et au final compréhensible, à partir de données.*

Comme l'explique ce dernier auteur, l'ECBD peut se décomposer en de nombreuses étapes plus ou moins complexes mais la figure 1.1 en donne une vision synthétique. Parmi les grandes étapes de l'ECBD, on peut distinguer :

- la **sélection** qui crée un ensemble de données à étudier ;
- le **prétraitement** qui vise à enlever le bruit et à définir une stratégie pour traiter les données manquantes ;
- la **transformation** où l'on recherche les meilleures structures pour représenter les données en fonction de la tâche ;
- la **fouille de données** : la fouille proprement dite et la définition de la tâche : classification, recherche de modèles... et la définition des paramètres appropriés ;
- **l'interprétation et l'évaluation** pendant laquelle les patrons extraits sont analysés. La connaissance qui en est ainsi extraite est alors stockée dans la base de connaissances.

Selon notre définition de la connaissance, la base de connaissances doit pouvoir être

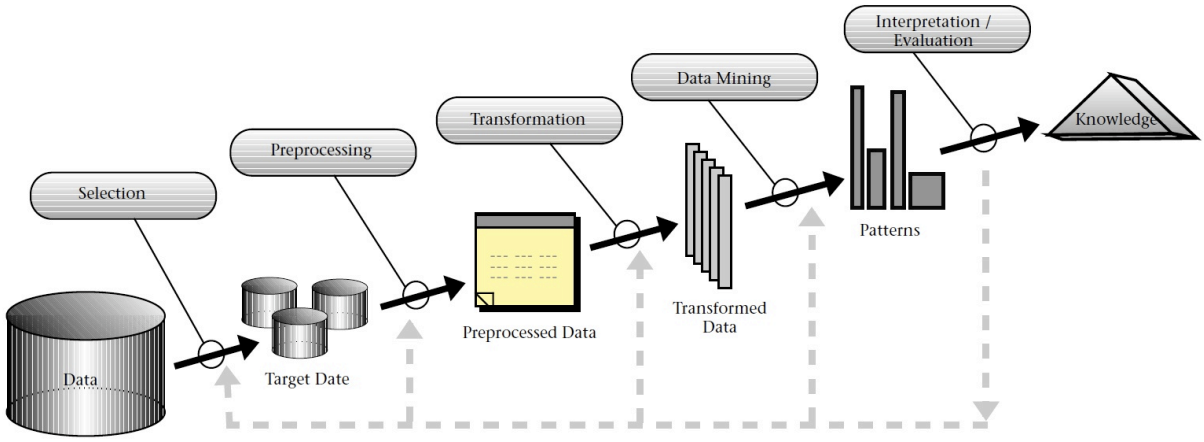


FIGURE 1.1 – Schéma global de l'ECDB d'après Fayyad[FPSS96]

utilisée pour raisonner, c'est-à-dire produire une nouvelle information. parmi les différents formalismes permettant de représenter les connaissances, nous nous intéressons plus particulièrement aux logiques de descriptions.

La fouille de données concerne donc spécifiquement les algorithmes permettant de faire émerger des données, des régularités ou des corrélations. Nous nous intéressons plus spécifiquement à deux types de méthodes symboliques : l'extraction de règles d'association [AIS93] et l'analyse formelle de concepts [GW99]. Un certain nombre d'expérimentations dans le domaine de la fouille de textes exploitent ces méthodes[BCM05, Mae02, SM01a]. Dans le domaine des textes où la dimension relationnelle des données est souvent forte, l'analyse relationnelle des concepts (RCA) [HNRHV07] est un nouveau cadre très prometteur (voir section 3.4.2.0). La RCA construit une famille de treillis à partir de la description de familles d'objets. Les concepts formels d'un treillis sont alors reliés aux concepts formels d'un autre treillis par les relations du domaine.

La construction de connaissances à partir de textes et le raisonnement sur ces connaissances est au coeur de mon projet de recherche.

1.2 Vers une définition de la fouille de textes

Contrairement au consensus qui a émergé autour du terme extraction de connaissances à partir de bases de données (ECBD), de son processus et notamment du positionnement de la fouille de données comme une de ses étapes (voir Figure 1.1 en section 1.1.4), le terme de fouille de textes a été assez différemment utilisé et n'apparaît pas, dans la littérature, comme une étape dans le processus d'extraction de connaissances : [Roc04], par exemple, applique des méthodes de fouilles de données à des textes au niveau de l'extraction d'information. Le Défi en Fouille de Textes (DeFT) est un atelier qui se tient chaque année depuis 2005, dont les objectifs ne portent pas sur l'extraction de connaissances à partir de textes. En revanche, [Che04] désigne le processus complet du texte à la production de pépites de connaissances. *A posteriori* et au vu de mon expérience

de construction d'ontologie à partir de textes dans divers domaines, il me semble cohérent de considérer que la fouille de textes englobe tout ce schéma.

De même que pour les données, les connaissances ne se trouvent pas dans les textes de façon intrinsèque ; un simple écrémage des termes les plus fréquents ou une classification ne suffisent pas à eux seuls. En revanche, de façon similaire à la terminologie qui fait émerger des termes d'un corpus (voir section 1.1.3.0), les connaissances sont le résultat d'une construction qui commence dès la collecte du corpus et d'autres formes de ressources existantes, qui s'appuie sur les textes, les mots, la mise en forme, mais qui fait appel de façon très entremêlée au traitement automatique de la langue, à la fouille de données, à l'interaction et la validation de l'expert. Le processus n'est donc pas aussi linéaire que la figure 1.1 le laisse penser, et les boucles de retour vers les étapes antérieures jouent un rôle essentiel. Il faut concevoir ces étapes comme des processus concourants.

Les applications nécessitant des connaissances sont de plus en plus nombreuses et dans des domaines diversifiés et les textes sont souvent un vecteur important d'information. S'il est impossible de construire l'Ontologie d'un domaine et encore moins une ontologie universelle, il est, en revanche, essentiel de disposer d'outils permettant de construire les ontologies, de les adapter à un nouveau problème, mais également de les enrichir et de les maintenir à jour, comme le souligne également [Smi08] : "Scientific ontologies are often highly complex. They are subject to a high velocity of change, not only in virtue of scientific advance, but also because the associated computational technologies are themselves rapidly evolving. As new applications for ontology-based technology are identified, so new ontologies are being created (...)"

L'objectif que nous nous donnons au travers de la fouille de textes est de permettre à un expert ou un ensemble d'experts d'un domaine d'avoir une vision synthétique de son domaine, et notamment des dernières avancées en les représentant dans un langage de représentation de connaissances.

Définition 1 *La fouille de textes est un ensemble de processus permettant, à partir d'un ensemble de ressources textuelles, de construire des connaissances pouvant être représentées dans un langage formel de représentation de connaissances et exploitées pour raisonner sur le contenu des textes.*

Je propose donc de décomposer la fouille de textes en 7 étapes qui pourraient, bien évidemment, être décomposées ou affinées :

1. Sélection des ressources
2. Prétraitements
3. Extraction d'informations
4. Fouille de données
5. Interprétation/validation
6. Représentation formelle
7. Annotation, raisonnement

La **sélection des ressources** sur un domaine, bien que souvent assez opportuniste, joue un rôle déterminant dans l'adéquation de l'ontologie construite à la tâche pour laquelle elle est construite. Certains textes apportent une connaissance de nature plutôt

ontologique (connaissance consensuelle), alors que d'autres relatent une expérimentation dont les résultats pourraient être remis en cause par d'autres expérimentations (conditions différentes, expérience non reproductible...). Certaines ressources sont centrales par rapport à la tâche et peuvent être constituées, par exemple, par un ensemble de résumés d'articles scientifiques sur les phénomènes de mutation génétique des bactéries. D'autres ressources sont plutôt de nature complémentaire et introduisent des connaissances déjà acquises dans un autre contexte. Il s'agit par exemple de prendre en compte les familles des bactéries. Ces connaissances ou informations peuvent se trouver dans des thésaurus, des ontologies, des textes didactiques. Ce type de ressources n'en est pas moins essentiel à la qualité de l'ontologie produite.

Enfin, les ressources doivent pouvoir être complétées au cours des différentes itérations du processus de fouille, pour pallier les manques identifiés par les experts lors de la validation des modèles extraits des données.

Les techniques impliquées dans cette étape sont issues de la recherche d'information : requête sur des bases documentaires, classification ou catégorisation des réponses. La difficulté consiste à réduire au mieux la dispersion (éviter des textes sans intérêt pour l'ontologie à construire) tout en préservant une bonne couverture des phénomènes.

Les **prétraitements** sont de natures très différentes. Cela peut concerner des questions d'encodage de caractère, de structure de documents, d'étiquetage morpho-syntaxique... Les outils généralement utilisés dans la suite du processus de fouille de textes travaillent sur du texte seul, sans mise en forme (ou une mise en forme très réduite telle que l'italique, la graisse...). Il est difficile de délimiter ce qui concerne les prétraitements ou l'extraction d'information dans l'absolu. Ainsi, la segmentation d'un texte peut extraire la partie pertinente d'un texte et constituer un prétraitement ou bien permettre de découper un texte en plusieurs sous-parties et pourrait alors être assimilée aux premières étapes d'extraction d'information.

L'**extraction d'information** est un terme générique ici qui permet de passer d'une forme textuelle à des données structurées. Cette étape repose, dans les travaux actuels, sur des outils issus du TAL : les analyseurs syntaxiques tentent une analyse complète des énoncés alors que les méthodes d'EI précédemment citées (voir section 4.3) procèdent en plusieurs "couches" de traitements locaux.

La **fouille de données** fait émerger des données extraites des textes des modèles qui, une fois validés par les experts, formeront des "unités de connaissance". C'est une étape essentiellement "algorithmique" faisant émerger des données des associations, des règles ou des classes selon la méthode utilisée.

L'**interprétation/validation** est menée par un ou plusieurs experts du domaine qui doivent également connaître la tâche pour laquelle ce travail d'extraction de connaissances est réalisé. De ce fait, l'évaluation est un problème difficile. Du point de vue du domaine, trouver un consensus entre experts sur l'organisation des connaissances du domaine demande beaucoup de temps. Au final, le résultat peut être très distant des modèles proposés par le processus de fouille de données ce qui pose problème, notamment lors de l'ajout de nouvelles données. Du point de vue de la tâche, il est souvent difficile de prévoir l'impact de telle ou telle décision sur les raisonnements qui seront opérés sur les connaissances.

La **Représentation formelle :** les unités de connaissances validées par les experts sont alors codées dans un langage formel de représentation de connaissances pour consti-

tuer l'ontologie. Nous nous intéressons plus particulièrement aux logiques de descriptions pour l'interrogation de l'ontologie avec des langages comme SPARQL ou pour raisonner.

L'**Annotation et le raisonnement :** Dans mon projet de recherche, l'annotation sémantique fait partie du raisonnement sur les textes dans le sens où l'annotation est construite de façon progressive, avec des connaissances de plus en plus riches. Maintenir la cohérence entre les connaissances et les textes annotés se heurte au fait que la trace linguistique d'un concept peut être complexe et qu'un certain nombre d'outils d'annotation qui se rapprochent d'une indexation terminologique (la recherche du terme associé à un concept), ne fonctionnent plus. D'autre part, la trace dans les textes n'est souvent que partielle, des connaissances implicites (et donc externes aux textes) devant être introduites pour la construction de la base de connaissances et la description des concepts. Quant au raisonnement, il est probablement la raison de vivre des bases de connaissances. Un raisonnement erroné dans un système à base de connaissances doit impliquer des modifications ou des enrichissements de la base de connaissances, avec, si possible, un contrôle de la non-régression du système (des réponses du système, précédemment correctes qui ne le seraient plus après modification). Cette dernière étape est donc bien indissociable du processus d'extraction de connaissances.

1.3 Textes versus données : quelques exemples

1.3.1 Qu'extraire des textes ?

Il existe des différences significatives entre l'extraction de connaissances à partir de données issues de bases de données et l'extraction à partir de données textuelles. Un processus d'extraction de connaissances qui prend en entrée ses données dans une base de données dissimule une étape essentielle : celle de la construction du modèle des données, étape qui conduit à définir les informations pertinentes pour un domaine et à la définition d'une certaine structure pour la base. Cela ne supprime pas pour autant les étapes de sélection, de prétraitement ou de transformation pour traiter des données complexes comme les intervalles ou les graphes pour lesquelles un échelonnage des données pour obtenir des données binaires est parfois nécessaire. Pourtant, un certain nombre d'imprécisions ou d'ambiguïtés ont déjà été levées. Lorsque l'on parle de textes, il faut d'abord souligner que la dimension textuelle est souvent assez pauvre car la plupart des expériences en fouille de textes portent généralement sur des résumés, par exemple, des résumés d'articles scientifiques, qui sont des textes courts et faiblement structurés. Mais, contrairement aux données, les textes regorgent d'imprécisions, d'ambiguïté et l'information utile est parfois dissimulée dans des tournures complexes. Je donne ci-dessous quelques exemples pour illustrer cette complexité liée à la langue.

Le prétraitement des textes est souvent vu comme une étape d'identification des "objets", de leurs propriétés et des relations avec d'autres objets du domaine et repose souvent sur des outils d'extraction d'information (voir section 4.3). Ainsi, la phrase *"The transcriptional $NF-_KB$ activity was increased by the overexpression of the tissue inhibitors of metalloproteinase TIMP-2."* exprime une relation de dépendance entre deux phénomènes, l'activité $NF-_KB$ et la surexpression de la protéine TIMP-2. Cependant, il existe d'autres

L'effet de l'addition d'enzymes pectolytiques au moût sur l'évolution de la fermentation ainsi que sur la composition et la qualité du cidre a été étudié. La clarification enzymatique améliore la couleur du cidre et les qualités sensorielles du cidre.

FIGURE 1.2 – Exemple de résumé en agro-alimentaire

relations "cachées". Le terme "tissue inhibitors of metalloproteinase" (inhibiteur tissulaire des métalloprotéinases) peut-être vu comme un terme figé ou analysé selon des principes de morphologie constructionnelle comme étant une protéine dont le rôle est de protéger les tissus de l'action de métalloprotéïnases, exprimant ainsi une relation de type "protection". De la "granularité" de l'analyse dépend donc l'information qui sera extraite des textes.

La figure 1.2 montre que le niveau de la phrase n'est parfois pas suffisant et que le niveau du texte (plus exactement, du discours) doit être pris en compte. La première phrase (on notera au passage la complexité syntaxique de la phrase) ne donne que l'objectif de l'étude, sans en préciser les résultats. La seconde donne les résultats (de façon assez vague) de l'expérience. Cette exemple illustre également à quel point certains phénomènes langagiers sont complexes à traiter si l'on veut expliciter toute l'information contenue dans un énoncé. Dans cet exemple, on apprend que la "clarification enzymatique du cidre" est, en fait, une "addition d'enzymes pectolytiques" et que "l'évolution de la fermentation et la composition du cidre" se manifeste par la "couleur et les qualités sensorielles du cidre".

La figure 1.3 est le résumé d'une étude clinique sur l'effet du dosage et du nombre de prises de l'érythropoïétine. Dans l'expérimentation relatée dans ce texte, deux groupes de patients ont été constitués, groupes sur lesquels ont été expérimentés différents modes d'administration (voir les parties mises en gras dans le texte). Les résultats sont alors donnés pour chacun des groupes : certaines corrélations n'ont pu être mises en évidence – *"There was no statistically significant correlation between change in hemoglobin level and tumor response for either group"* – et d'autres ont pu l'être – *"there was not correlation between hemoglobin change and serum cytokine changes from baseline, except for IL-6 in Group A"*. La difficulté réside ici dans la capacité à extraire du texte les caractéristiques des patients de chacun des groupes et leur mise en relation avec les résultats obtenus. Mais il faut également déterminer quelle information va être construite à partir de ces textes. Faut-il, par exemple, créer des instances de patients et leur attribuer les caractéristiques liées à leur groupe ou est-ce une information plus globale avec un nombre de patients, des conditions expérimentales et des résultats.

Ces quelques exemples doivent permettre de se convaincre que les données textuelles sont très différentes des données issues de bases de données ou de données, comme des mesures, et qu'il ne suffit pas de repérer des mots ici ou là pour remplir des champs qui alimenteraient un processus de fouille.

1.3.2 Quelles connaissances et pour quoi faire ?

Comme nous l'avons déjà souligné, une connaissance est construite sur un domaine particulier et en fonction d'une tâche à réaliser. Voici quelques exemples de questions que

Assessment of the efficacy of two dosages and schedules of human recombinant erythropoietin in the prevention and correction of cisplatin-induced anemia in cancer patients.

Despite the numerous studies demonstrating the effectiveness of epoetin a (human recombinant erythropoietin) versus placebo in cisplatin-induced anemia of cancer patients, data are lacking on the most effective doses and schedules of administration of epoetin a in this setting. The aim of the present study was to assess the best dose and schedule of administration of epoetin a in cancer patients with cisplatin-induced anemia. This was an open, randomized, single-institution phase II study comparing the ability of two doses and schedules of epoetin a of preventing and/or correcting anemia (. . .) **The eligible patients were randomly assigned to treatment with either : a) subcutaneous epoetin a 150 U/kg three times a week for up to 12 consecutive weeks (Group A) ; b) subcutaneous epoetin a 50 U/kg daily for up to 12 consecutive weeks (Group B)**. The following laboratory parameters were assessed before the study entry and during the study) : hemoglobin (weekly) ; serum iron, transferrin and ferritin (before entry). The following immunological parameters were assessed before and after study end : Interleukin (IL)-1a, IL-1 , IL-6 and Tumor Necrosis Factor (TNF) a. Twenty patients were enrolled, data were available for 17. **Nine patients were assigned to Group A and 8 to Group B**. **No statistically significant difference** of hemoglobin level was found between the 2 groups at baseline, at month 1, 2 and 3, neither in the comparison of the change from baseline between the two groups. **In Group A** fewer transfusions were administered per patient per month after the first month of epoetin a therapy, **compared to Group B**. No significant difference was found as for transfusion requirements at month 1, 2 and 3 **between Group A and B**. The epoetin a dose administered was slightly higher than that projected. Epoetin a was well-tolerated. There was no statistically significant correlation between change in hemoglobin level and tumor response **for either group**, neither between change in hemoglobin level and change in ECOG score from baseline to final was observed. The changes from baseline of IL-1a and IL-1 , IL-6 and TNFa were not remarkable nor univocal **in either group**, there was not correlation between hemoglobin change and serum cytokine changes from baseline, except for IL-6 **in Group A.**

FIGURE 1.3 – Exemple de résumé en pharmacologie

des experts en astronomie ou en microbiologie sont venus nous soumettre et que nous avons abordées, notamment dans les thèses de Hacène Cherfi et de Rokia Bendaoud.

En astronomie, les astronomes souhaitent classer les objets célestes en fonction de leurs propriétés mais cherchent également à identifier les propriétés monothétiques de ces objets, ou plus généralement des propriétés discriminantes, permettant de classer tel objet dans telle classe. Il s'agit donc d'identifier dans les textes des propriétés définitoires des classes d'objets. Nous avons pu montrer, par exemple, qu'à partir des articles scientifiques en astronomie, une nouvelle classe d'objets célestes `Eclipsing star` pourrait être définie comme celle des étoiles qui émettent et qui s'éclipsent. Cette classe regrouperait alors des objets célestes qui, actuellement, sont regroupés avec d'autres sous le nom d'étoile. Les objets célestes concernés sont les suivants `Algol, SAO 186497, HS Her, TW Cnc, V649 Cas, MM Herculis` et `Y Cygni`.

En microbiologie, les phénomènes de résistance des bactéries aux antibiotiques ont fait l'objet de nombreuses publications. L'idée qui a guidé nos travaux était de construire une représentation synthétique des connaissances sur ces phénomènes : regrouper dans une même classe les entités(antibiotiques, bactéries, gènes...) qui fonctionnent de façons identiques, et expliquer les phénomènes de résistance par la mise en évidence de relations entre ces classes.

Dans un domaine qui nous est plus communément accessible et probablement un peu plus intuitif comme la cuisine, la fouille de textes construit des connaissances permettant de raisonner sur un ensemble de recettes et apporte des réponses à des questions pratiques de tous les jours : est-ce que je peux remplacer les pommes par des poires dans une tarte ? Comment dois-je adapter la recette et faut-il préparer les poires différemment des pommes ? La construction d'une ontologie structurant les ingrédients permet de déterminer des ingrédients "proches" et peut bien répondre à de telles questions. Mais cela ne suffit pas, il faut également identifier les différents modes de préparation d'un ingrédient pour que la substitution d'un ingrédient se fasse selon des usages culinaires "reconnus" ou acceptables. Enfin, pour se rapprocher de la problématique de l'astronomie, on peut souhaiter construire des classes de plats, par exemple, les soupes ou les ragoûts, et tenter de définir ces classes par la présence de certains ingrédients ou de certains modes de préparation.

1.4 Projet de recherche sur la fouille de textes

1.4.1 Une recherche articulée autour des méthodes de fouille de données à base de motifs

Mon projet de recherche vise à raisonner sur le contenu des textes en exploitant des méthodes symboliques de fouille de données : l'extraction de règles d'association et l'analyse formelle de concepts. Je montrerai plusieurs facettes de l'utilisation de ces méthodes pour les textes : la construction d'ontologies comme synthèse des avancées scientifiques dans un domaine, la construction d'ontologies en vue de leur exploitation par un moteur de raisonnement (par exemple de raisonnement à partir de cas), mais aussi l'adaptation ou la transformation de textes exploitant les treillis ou l'extraction d'information par ex-

traction de motifs. Les deux chapitres suivants (Chapitres 2 et 3) sont consacrés à la présentation des règles d'association et des treillis dans cette optique.

Le chapitre 4 relie ces travaux à des travaux issus de domaines connexes à ceux de la fouille de données comme la recherche d'information ou le traitement automatique de la langue, questions incontournables dès lors que nous partons de textes réels.

Enfin, le chapitre 5 formule mon projet de recherche, propose quelques cadres applicatifs sur lesquels j'ai travaillé ou souhaiterais travailler et j'expliciterai quelques uns des défis posés, notamment au travers de problèmes pouvant faire l'objet de sujets de thèse.

Ce document se termine par **trois articles scientifiques placés en annexe** que j'ai co-publiés et qui illustrent bien les orientations de mes travaux et les questions que je souhaite aborder dans mon projet de recherche. Le présent document est donc conçu comme un complément à ces articles, pour, notamment, les replacer dans un contexte scientifique un peu plus large. Je suggère donc au lecteur de lire ces articles au moment où j'y fais référence. Le premier article est un chapitre de livre [CNT09] qui porte sur la définition d'un indice de vraisemblance pour des règles d'association extraites de textes et qui détermine à quel point une règle reflète une connaissance déjà acquise. Le second article vient d'être accepté à la conférence PKDD 2010 et porte sur la fouille de données issues de la pharmacovigilance. Bien que portant sur des données et plus que des données textuelles, cet article illustre les travaux que je mène sur la FCA comme méthode de fouille de données. Il constitue un point de départ pour faire de la pharmacovigilance un processus de fouille de données guidé par des connaissances du domaine. Enfin, le troisième article [BNT08b] a été publié à la conférence EKAW et illustre l'utilisation de la FCA pour la construction d'ontologie à partir de textes.

J'invite donc le lecteur à lire ces articles soit maintenant, en préalable aux sections suivantes, soit au moment où je les cite comme exemple de mes travaux dans ces sections.

1.4.2 Un nouveau système de construction d'ontologie ?

Si je désigne par ontologie les connaissances capitalisées sur tel ou tel domaine d'expérimentation comme par exemple, en astronomie ou en microbiologie, le développement d'un environnement de construction d'ontologies est un travail complexe, en partie d'ingénierie, et les problèmes qui se posent dépassent les travaux que je présentent ici ; ils font appel à des compétences diverses, depuis l'ergonomie jusqu'à la maîtrise des formalismes de représentation des connaissances en passant notamment par la linguistique. Deux projets me semblent bien représentatifs des questions que soulèvent le développement de tels systèmes.

Terminae [BS99] (http ://www.springerlink.com/content/5fkmhc4wpq0nuqd5/) fait partie des premiers projets àaller au delà des méthodologies existantes et à construire, à la fois théoriquement et concrètement, un environnement pour la construction d'ontologies. Il tire profit des travaux menés par le groupe *Terminologie et Intelligence Artificielle* créé en 1993 et qui a introduit la notion de base de connaissances terminologique qui établit le lien entre la description linguistique d'un terme et le concept qui lui est associé. L'originalité de Terminae est de proposer une formalisation "progressive" des concepts, en trois étapes : (1) une définition de la notion par ses relations avec d'autres entrées lexicales, (2) une définition "plus formelle" dans laquelle les relations sont traduites en primitives issues

d'une liste de relations prédéfinies, (3) une définition formelle en logique de descriptions. Le projet ANR Dafoe [SCAG+10], qui vient de se terminer, se place dans le même esprit que Terminae. Il vise à faire coopérer un ensemble d'outils pour réaliser une plateforme technique pour concevoir une ontologie. Le cadre méthodologique est constitué de quatre étapes, par ailleurs partagées par la plupart des méthodes de construction d'ontologies : constitution d'un corpus de documents, analyse linguistique du corpus, conceptualisation, opérationnalisation de l'ontologie.

Text2Onto [CV05] est également un environnement pour la construction d'ontologies à partir de textes. Les points forts mis en avant par les auteurs de cette plateforme sont (1) dans la représentation des connaissances à un méta niveau – représentation qui peut être traduite dans un formalisme de représentation de connaissances mais qui associe également aux connaissances un degré de confiance – (2) une interaction forte avec l'utilisateur final, (3) un suivi des changements dans les données qui induisent un changement dans l'ontologie.

Mon projet de recherche, vis-à-vis de ces expériences en construction d'ontologie peut apparaître comme quelque peu idéaliste et peut-être très limité . Il s'agit en effet pour moi de tirer profit au maximum des méthodes formelles de fouille de données dont je dispose pour prendre en compte des objets complexes, syntaxiques (arbres, relations de dépendances) ou sémantiques (frames, graphes...), d'exploiter les environnement sémantiques et coopératifs récents comme, par exemple les wikis sémantiques, et ainsi affiner la traçabilité entre informations et connaissances en travaillant en parallèle sur la construction de l'ontologie et l'annotation sémantique de textes. De telles méthodes et outils peuvent constituer des réponses à certaines questions encore mal résolues dans les environnements de construction d'ontologie et ainsi s'intégrer à ces projets plus vastes à différents niveaux, pour la constitution de ressources lexicales, pour l'extraction d'information ou pour l'analyse syntaxique ou encore pour la phase de conceptualisation.

Chapitre 2

Les règles d'association

Les règles d'association ont été étudiées en analyse de données [GD86, Lux91], puis en fouille de données afin de trouver des régularités, des corrélations dans des bases de données de grandes tailles [AIS93]. Les règles d'association comme méthode de fouille de données ont un aspect séduisant par l'apparente facilité à interpréter les résultats. Ainsi, on peut reprendre l'exemple courant du panier de la ménagère où les règles d'association montreraient que lorsqu'un homme achète des couches-culottes dans un hypermarché, il achète également de la bière. Il est donc possible d'identifier quels sont les produits achetés en même temps que d'autres produits et de prévoir, en conséquence, une disposition particulière des produits ou des offres promotionnelles liées à la consommation des clients. Un certain nombre d'expériences ont également porté sur la fouille des fichiers de log sur des sites internet pour identifier quelles sont les pages web accédées "autour" d'une page donnée. Les règles d'association ont également été appliquées aux textes. Dans les premiers travaux, [FD95b, FFK$^+$98] identifient des relations de corrélation entre les mots ou entre les termes. Mais, des travaux plus avancés s'appuyant sur des prétraitements plus sophistiqués tant au niveau linguistique qu'au niveau de la structure des documents, se sont, par exemple, intéressés à détecter des fautes de typographie dans des documents [LC04].

En première approximation, on peut définir une règle d'association comme une règle de la forme $\texttt{B} \longrightarrow \texttt{H}$ dans laquelle `B` et `H` sont des ensembles d'attributs. Une telle règle peut être interprétée de la façon suivante : "Les individus qui possèdent les attributs de `B` possèdent également les attributs de `H`". `B` et `H` sont appelés des motifs. Ainsi, l'exemple du paragraphe précédent se traduirait par la règle $\texttt{homme}, \texttt{couche\_culotte} \longrightarrow \texttt{bière}$. Il existe de nombreux algorithmes d'extraction de règles d'association, qui, généralement extraient en premier lieu des motifs fréquents puis ensuite les règles d'association.

2.1 Contexte formel

Nous évoquions dans le chapitre 1, la question de la représentation d'un document. Si la recherche d'information privilégie une représentation vectorielle d'un document par un vecteur de mots-clés ou de termes pondérés par le TF-IDF (Term Frequency - Inverse Document Frequency), nous lui préférons une représentation booléenne dans laquelle les documents sont représentés par un ensemble de mots, par exemple, ou encore des des-

TABLE 2.1 – Exemple d'un contexte formel où $\{\mathtt{a}, \mathtt{b}, \ldots\}$ sont des termes et $\{\mathtt{d}_1, \mathtt{d}_2, \ldots\}$ sont des textes.

$\mathcal{I}$	a	b	c	d	e
$\mathtt{d}_1$	0	1	1	0	1
$\mathtt{d}_2$	1	0	1	1	0
$\mathtt{d}_3$	1	1	1	1	0
$\mathtt{d}_4$	1	0	0	1	0
$\mathtt{d}_5$	1	1	1	1	0
$\mathtt{d}_6$	1	0	1	1	0

cripteurs plus complexes issus de prétraitements linguistiques. La sensibilité au bruit ou au silence de la représentation booléenne est bien connue, surtout dans le domaine documentaire où un mot peut être présent dans un document mais absent dans un autre, pourtant très proche thématiquement : la conséquence est donc que la présence ou l'absence d'un descripteur pour un document va changer la classe dans laquelle il sera placé. En revanche, la représentation booléenne est plus appropriée que la description vectorielle pour les approche sémantique – quand il s'agit de construire un sens – comme le souligne Barthes [Bar66] (voir section 1.1.2).

Les règles d'association et l'analyse formelle de concepts partagent un mode commun de représentation des données. Les données sont constituées d'objets décrits par des attributs. Les objets peuvent être appelés transactions dans le cas de l'extraction de règles d'association, ou documents dans les applications aux textes. Les attributs sont parfois appelés items ou descripteurs (dans le documentaire). L'ensemble des données est appelé contexte formel.

Définition 2 (Contexte formel) *Un contexte formel est un triplet* $\mathbb{K}=(\mathtt{G}, \mathtt{M}, \mathtt{I})$ *où* $\mathtt{G}$ *est un ensemble fini d'objets,* $\mathtt{M}$ *un ensemble fini d'attributs et* $\mathtt{G}$ *et* $\mathtt{M}$ *sont des ensembles disjoints.* $\mathtt{I} \subseteq \mathtt{G} \times \mathtt{M}$ *est une relation binaire entre* $\mathtt{G}$ *et* $\mathtt{M}$.

La figure 2.1 donne un exemple de contexte formel (repris de [Sza06]) pour lequel $\mathtt{G} = \{\mathtt{d}_1, \mathtt{d}_2, \mathtt{d}_3, \mathtt{d}_4, \mathtt{d}_5, \mathtt{d}_6\}$ est l'ensemble des objets et $\mathtt{G} = \{\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{d}, \mathtt{e}\}$ est l'ensemble des attributs. $\mathtt{I}$ est la relation qui associe à chaque objet un ensemble d'attributs (0 si l'objet ne possède pas l'attribut et 1, s'il le possède). Ainsi, d_1 possède les propriétés $\mathtt{b}$, $\mathtt{c}$, et $\mathtt{e}$.

2.2 Les motifs fréquents

Définition 3 *Étant donnés un ensemble d'objets* $\mathtt{G}$ *et un ensemble d'attributs* $\mathtt{M}$*, un* **motif** *est un sous-ensemble de* $\mathtt{M}$*. On dit qu'un objet contient un motif si l'objet contient chacun des attributs du motif. La longueur d'un motif est le nombre d'attributs de ce motif. L'image d'un motif est l'ensemble des objets possédant ce motif. Le* **support** *d'un motif est le nombre d'objets possédant ce motif et le motif est* **fréquent** *si son support est supérieur à un seuil donné* σ_s.

En reprenant l'exemple du contexte formel donné en table 2.1 et pour un seuil $\sigma_s = 3$, `{a}` est un motif fréquent de longueur 1 et de support 5, `{ac}` est de longueur 2 et de support 4. Le motif `{abc}`, de longueur 3, a un support de 2, il n'est donc pas fréquent, de même que `{abcde}` qui est de longueur 5 et de support 0. Le support est une fonction monotone décroissante par rapport à la longueur du motif.

Le nombre maximum possible de motifs est 2^n où n est le cardinal de M et il est donc essentiel de proposer des algorithmes efficaces pour calculer les motifs fréquents. L'algorithme A Priori [AIS93, AS94, MTV94, AMS+96] est un algorithme très connu et très utilisé fonctionnant par niveau : il calcule les motifs de longueur m à partir des motifs de longueur $m-1$. Il repose sur deux principes duals qui permettent de réduire les combinaisons de motifs à explorer :

1. tout sous-motif d'un motif fréquent est fréquent ;
2. tout sur-motif d'un motif non fréquent est non fréquent ;

Un algorithme à niveau recherche en premier lieu les motifs fréquents de longueur 1. Les motifs non fréquents sont abandonnés alors que les motifs fréquents sont combinés ensemble pour former des motifs plus longs dont on teste le support pour ne garder que les fréquents. Le processus est réitéré jusqu'à ce qu'aucun nouveau motif ne puisse être formé.

À partir du contexte formel donné en table 2.1 et en posant $\sigma_s = 2$, les motifs fréquents de longueur 1 sont `{a}`(3), `{b}`(5), `{c}`(5), `{d}`(5). Les motifs candidats de longueur 2 dont le support est testé sont `{ab}`(2), `{ac}`(4), `{ad}`(5), `{bc}`(3), `{bd}`(2), `{cd}`(4). `{e}` étant non fréquent, le motif `{ae}` n'est pas testé. Dans le cas présent, tous les motifs de longueur 2 sont gardés pour être combinés en motifs de longueur 3 dont le support sera testé. Et ainsi de suite.

Des optimisations de cet algorithme ont été proposées [PBTL99a, PBTL99b] pour gagner en efficacité dans le cas de très grands contextes formels. Ces optimisations exploitent la notion de motif fermé telle que définie en analyse formelle de concepts. Nous détaillerons cette notion de fermé dans le chapitre 3.

Les algorithmes d'extraction des motifs fréquents sont classés en trois catégories : les algorithmes à niveau, les algorithmes verticaux – profondeur d'abord– et les algorithmes hybrides – combinant les deux premières méthodes. [SN06] propose, au sein de la plateforme CORON développée dans l'équipe Orpailleur, trois nouveaux algorithmes pour l'extraction de motifs fréquents : **Zart** qui est une amélioration de l'algorithme *Pascal* [BTP+02] ; **Eclat-Z** est un algorithme hybride qui combine les algorithmes *Zart* – algorithme par niveau – et *Eclat* [ZPOL97] – algorithme vertical ; et **Charm-MFI** une extension de l'algorithme *Charm* [ZH02].

2.3 Les règles d'association

Définition 4 (Règle d'association) *Une* **règle d'association** *est une implication de la forme* $B \longrightarrow H$ *où* B *et* H *sont deux motifs. Le but de l'extraction de règles d'association est de déterminer si l'occurrence du motif* B *est associée à l'occurence du motif* H*.* B *est appelé prémisse et* H *conclusion de la règle. Le* **support** *de la règle est défini comme*

étant le support du motif $\mathtt{B} \sqcup \mathtt{H}$*, motif qui contient toutes les propriétés de* `B` *et de* `H`*. La* **confiance** *de la règle* $\mathtt{B} \longrightarrow \mathtt{H}$ *est définie par le rapport* $\mathtt{support}(\mathtt{B} \sqcup \mathtt{H})/\mathtt{support}(\mathtt{B})$*. La confiance peut être vue comme la probabilité conditionnelle de* `H` *sachant* `B` *soit* $P(H|B)$*.*

Si la confiance est supérieure ou égale à un seuil donné σ_c *et son support supérieur ou égal au seuil* σ_s*, la règle est dite* **valide**[1]*. Si la confiance est de 1, la règle est dite* **exacte** *(c'est une implication au sens logique), sinon, elle est approximative.*

La règle $\mathtt{B} \longrightarrow \mathtt{H}$ signifie que tout objet de `O` contenant le motif `B` contient aussi le motif `H` avec une probabilité égale à la confiance.

À partir de l'exemple de la Table 2.1 et avec $\sigma_s = 3$ et $\sigma_c = 3/5$, la règle $\mathtt{d} \longrightarrow \mathtt{ac}$ construite à partir du motif fréquent `acd` avec un support de 4 et une confiance de $\frac{4}{5} = 0,8$. De même, $\mathtt{c} \longrightarrow \mathtt{ad}$ construite à partir du même motif fréquent avec un support de 4 et une confiance de $\frac{4}{5} = 0,8$.

L'extraction des règles d'association se fait à partir des motifs fréquents. Si `M` est un motif fréquent, on commence par tester les règles dont la longueur de la conclusion est 1. Par exemple, `{a,c,d}` est un motif fréquent. $\mathtt{ac} \longrightarrow \mathtt{d}$ est testée et s'avère être une règle valide. Puis les règles ayant une conclusion de longueur 2 : $\mathtt{a} \longrightarrow \mathtt{cd}$ qui, elle aussi est valide. En revanche, si une règle $\mathtt{B} \longrightarrow \mathtt{H}$ est invalide à cause d'une confiance trop faible alors, toute règle $\mathtt{B} \longrightarrow \mathtt{H'}$ pour tout `H'` tel que $\mathtt{H} \subset \mathtt{H'}$ est invalide.

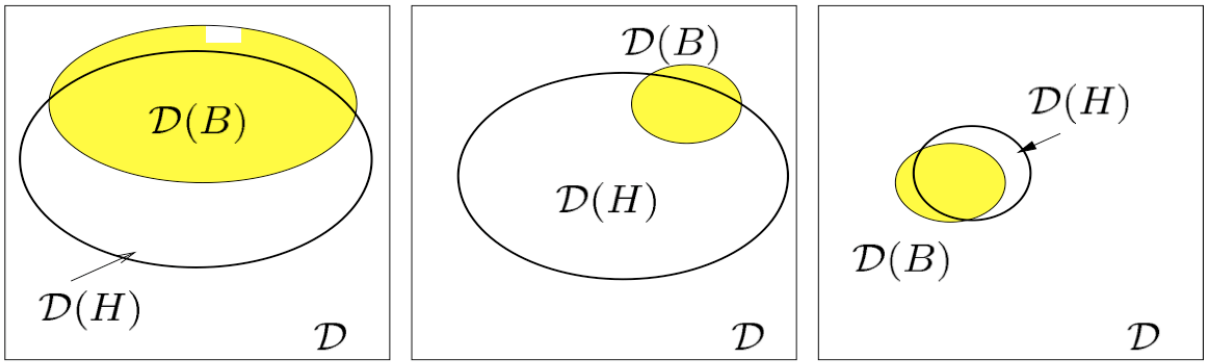
Le nombre de règles extraites est souvent très grand et peut atteindre plusieurs millions. Deux stratégies permettent de réduire ce nombre. La première est de ne générer que certains types de règles comme les bases génériques, les règles sur des motifs fermés ou les ensembles minimaux de règles d'association non redondantes. La seconde stratégie consiste à filtrer (ou ordonner) les règles extraites en se restreignant à des règles ayant une certaine forme (1 attribut en prémisse, par exemple) ou encore à leur associer des mesures de qualité et à les ordonner selon une ou plusieurs de ces mesures (voir section2.4).

2.4 Trier et filtrer les règles d'association

Une des solutions pour réduire le très grand nombre de règles extraites consiste à les trier ou les filtrer pour ne garder que les règles les plus "intéressantes". Comme nous l'avons souligné en section 1.1.4, tout élément de connaissance doit être validé par des experts mais leur temps et leur énergie sont comptés, il est donc très important de pouvoir concentrer leur interaction sur les points essentiels. Le support et la confiance sont deux mesures exploitées par les algorithmes d'extraction de règles d'association, notamment, pour en réduire le nombre, mais elles ne sont pas suffisantes pour qualifier l'intérêt de ces règles aux yeux d'un expert. Prenons l'exemple d'une règle $\mathtt{B} \longrightarrow \mathtt{H}$. Il est possible d'associer à cette règle des indices statistiques qui reflètent les différents cas de figure envisageables pour `B` et `H`.

[Fen07] et [GH07b] dressent un panorama des différentes mesures de qualité en fouille de données. Dans cet ouvrage, [GH07a] liste pas moins de 37 mesures différentes et montre clairement la complexité à en choisir une. Elles sont donc caractérisées selon 11 critères.

1. Ce terme "valide" utilisé en fouille de données ne doit pas être confondu avec le fait que l'implication (en tant que formule) soit valide ou non, du point de vue de la logique



(a) B et H fréquents **(b) B rare, H fréquent** **(c) B et H rares**

FIGURE 2.1 – Différentes situations illustrant une règle $\mathtt{H} \longrightarrow \mathtt{B}$

Je ne présente que quelques unes de ces mesures, parmi les plus utilisées, pour trier ou filtrer les règles d'association.

2.4.1 Mesures dites "objectives"

Le mesures objectives sont dites "dirigées par les données" car ces mesures ne reposent que sur les motifs et leur distribution dans les données.

Soit $\mathcal{D}(\mathtt{B})$, $\mathcal{D}(\mathtt{H})$ et $\mathcal{D}(\mathtt{B} \sqcup \mathtt{H}) = \mathcal{D}(\mathtt{B}) \cap \mathcal{D}(\mathtt{H})$ les sous-ensembles de textes de $\mathcal{D}$ possédant respectivement tous les termes de B, H et $\mathtt{B} \sqcup \mathtt{H}$ (*cf.* FIGURE 2.1). $\mathtt{B} \sqcup \mathtt{H}$ désigne le motif composé des des termes de B et des termes de H. Trois valeurs de probabilités ont un impact sur la valeur des mesures que nous utilisons : $\mathtt{P}(\mathtt{B})$, $\mathtt{P}(\mathtt{H})$ et $\mathtt{P}(\mathtt{B} \sqcup \mathtt{H})$ qui se définissent par la formule générale suivante : $\left(\mathtt{P}(\mathtt{X}) = \frac{|\mathcal{D}(\mathtt{X})|}{|\mathcal{D}|}\right)$ compris entre 0 et 1. $\mathtt{P}(\mathtt{B} \sqcup \mathtt{H})$ est le support de la règle. La probabilité conditionnelle $\mathtt{P}(\mathtt{H}|\mathtt{B}) = \frac{\mathtt{P}(\mathtt{B} \sqcup \mathtt{H})}{\mathtt{P}(\mathtt{B})}$ en est la confiance.

Plus $\mathcal{D}(\mathtt{X})$ est grand et couvre l'ensemble $\mathcal{D}$, plus X est fréquent et plus $\mathtt{P}(\mathtt{X})$ est fort et donc proche de 1. Si la règle est constituée de motifs B et H très fréquents, alors ces motifs sont partagés par presque tous les textes. Par conséquent, le volume de connaissances convoyé par ces motifs, du point de vue de la découverte de connaissances, est faible ou nul.

- Pour le cas (a) de la Figure FIGURE 2.1, $\mathtt{P}(\mathtt{B})$ et $\mathtt{P}(\mathtt{H})$ sont toutes deux proches de 1, les règles associées sont considérées comme peu informatives. Un ensemble de termes présent dans presque tous les textes implique, très probablement, un autre ensemble présent dans tous les textes. Il y a de grandes chances que les termes de B et H soient des termes génériques du domaine. Par exemple, la règle `mutation` $\longrightarrow$ `résistance` ne présente pas d'intérêt si `mutation` et `résistance` sont deux termes très répandus qui ont permis de sélectionner les textes du corpus d'expérience ;
- Le cas (b), lorsque $\mathtt{P}(\mathtt{B})$ est plus faible, paraît, en ce sens, plus intéressant. L'inconvénient est que tout texte qui possède B aura tendance à posséder H ;
- Le cas (c) est le plus intéressant. Les termes de $\mathtt{P}(\mathtt{B})$ et $\mathtt{P}(\mathtt{H})$ y sont rares et appa-

raissent presque à chaque fois ensemble ($\mathtt{P}(\mathtt{B} \sqcup \mathtt{H}) \simeq \mathtt{P}(\mathtt{B}) \simeq \mathtt{P}(\mathtt{H})$). Ces termes sont donc vraisemblablement reliés dans un contexte du domaine ;
– Le quatrième cas possible (B fréquent, H rare) n'est pas considéré ici. Ce cas correspond à un seuil de *confiance* faible $\left(\frac{\mathtt{P}(\mathtt{B} \sqcup \mathtt{H})}{\mathtt{P}(\mathtt{B})} \ll 1\right)$.

Le support et la confiance

Les mesures de support et de confiance ne différencient pas complètement les cas (a), (b) et (c) de FIGURE 2.1. Le support représente l'intersection $\mathcal{D}(\mathtt{B}) \cap \mathcal{D}(\mathtt{H})$ et peut distinguer (a) de (b) et (c). La confiance représente l'inclusion de $\mathcal{D}(\mathtt{B})$ dans $\mathcal{D}(\mathtt{H})$ et n'est pas un facteur discriminant de ces trois cas. Pour ces raisons, les mesures de support et de confiance ne sont pas suffisantes pour identifier les cas : du plus significatif (c) vers le moins significatif (a). La suite du paragraphe montre que d'autres mesures statistiques de qualité sont capables de différencier les trois cas possibles de la FIGURE 2.1.

l'Intérêt ou le lift

L'**intérêt** (ou *lift*) mesure la déviation du support de la règle par rapport au cas d'indépendance. Rappelons que pour deux événements indépendants B et H, $\mathtt{P}(\mathtt{H}|\mathtt{B}) = \mathtt{P}(\mathtt{H})$ et donc $\mathtt{P}(\mathtt{B} \sqcup \mathtt{H}) = \mathtt{P}(\mathtt{B}) \times \mathtt{P}(\mathtt{H})$. La valeur de l'intérêt est donnée par :

$$\mathtt{int}\ [\mathtt{B} \longrightarrow \mathtt{H}] = \frac{\mathtt{P}(\mathtt{B} \sqcup \mathtt{H})}{\mathtt{P}(\mathtt{B}) \times \mathtt{P}(\mathtt{H})} \tag{2.1}$$

L'intérêt varie dans l'intervalle $[0, +\infty[$. Si B et H sont indépendants alors $\mathtt{int}\ [\mathtt{B} \longrightarrow \mathtt{H}] = 1$. Plus B et H sont incompatibles, plus $\mathtt{P}(\mathtt{B} \sqcup \mathtt{H})$ tend vers 0 et donc l'intérêt est proche de 0. Plus B et H sont dépendants, plus l'intérêt est supérieur à 1. Si l'intérêt a une valeur de 3, alors les exemples vérifiant la règle $\mathtt{B} \longrightarrow \mathtt{H}$ (ou la règle $\mathtt{H} \longrightarrow \mathtt{B}$ puisque ce sont les mêmes exemples et que l'intérêt est symétrique) sont trois fois plus nombreux que dans le cas de l'indépendance.

Puisque $\mathcal{D}(\mathtt{B}) \cap \mathcal{D}(\mathtt{H}) \subseteq \mathcal{D}(\mathtt{B})$ et $\mathcal{D}(\mathtt{B}) \cap \mathcal{D}(\mathtt{H}) \subseteq \mathcal{D}(\mathtt{H})$, plus $\mathcal{D}(\mathtt{B})$ et $\mathcal{D}(\mathtt{H})$ sont petits dans $\mathcal{D}$, plus la valeur de l'intérêt augmente. Si $\mathtt{P}(\mathtt{B} \sqcup \mathtt{H}) \simeq \mathtt{P}(\mathtt{B})$ alors $\mathtt{int}\ [\mathtt{B} \longrightarrow \mathtt{H}] \simeq \frac{\mathtt{P}(\mathtt{B})}{\mathtt{P}(\mathtt{B}) \times \mathtt{P}(\mathtt{H})} = \frac{1}{\mathtt{P}(\mathtt{H})}$, de la même manière $\mathtt{int}\ [\mathtt{B} \longrightarrow \mathtt{H}] = \frac{1}{\mathtt{P}(\mathtt{B})}$. Ainsi, quand $\mathtt{P}(\mathtt{B})$ ou $\mathtt{P}(\mathtt{H})$ tendent vers 0, l'intérêt tend vers $+\infty$. Par conséquent, les règles qui se trouvent dans le cas (c) sont classées en début. L'intérêt est symétrique : $\mathtt{int}\ [\mathtt{B} \longrightarrow \mathtt{H}] = \mathtt{int}\ [\mathtt{H} \longrightarrow \mathtt{B}]$.

La conviction

Comme le souligne [BMUT97], l'intérêt est surtout utilisé pour mesurer la dépendance entre B et H et, comme c'est une mesure symétrique, l'intérêt mesure essentiellement la cooccurrence de B et de H. Pour mesurer l'implication, [BMUT97] introduit donc la **conviction**.

$$\mathtt{conv}\ [\mathtt{B} \longrightarrow \mathtt{H}] = \frac{\mathtt{P}(\mathtt{B}) \times \mathtt{P}(\neg\mathtt{H})}{\mathtt{P}(\mathtt{B} \sqcup \neg\mathtt{H})} \tag{2.2}$$

La **conviction** est similaire à l'intérêt mais appliqué aux contre-exemples. Dans notre contexte, un contre-exemple correspond au motif $\mathtt{B} \sqcup \neg\mathtt{H}$ tel que $\neg\mathtt{H}$ signifie l'absence d'au moins un terme du motif $\mathtt{H}$. $|\mathcal{D}(\neg\mathtt{H})| = |\mathcal{D}| - |\mathcal{D}(\mathtt{H})|$ et $\mathtt{P}(\neg\mathtt{H}) = 1 - \mathtt{P}(\mathtt{H})$.

La conviction vaut donc $\left(\frac{1}{\mathtt{int}[\mathtt{B} \longrightarrow \neg\mathtt{H}]}\right)$; elle varie dans l'intervalle $[0, +\infty[$ et n'est pas symétrique. Comme l'intérêt, la conviction vaut 1 lorsque $\mathtt{B}$ et $\mathtt{H}$ sont indépendants. Elle n'est pas calculable pour les règles exactes (quand la confiance est à 100%) puisque $\mathtt{P}(\mathtt{B} \sqcup \neg\mathtt{H})$ vaut 0. Une conviction de 2 signifie que le nombre de contre-exemples de la règle est deux fois moins grand que celui attendu sous l'indépendance de $\mathtt{B}$ et de $\mathtt{H}$ [LT04]. La valeur de conviction augmente lorsque $\mathtt{P}(\neg\mathtt{H})$ est élevé ($\mathtt{P}(\mathtt{H})$ faible), $\mathtt{P}(\mathtt{B})$ est élevé et lorsque $\mathtt{P}(\mathtt{B} \sqcup \mathtt{H}) \simeq \mathtt{P}(\mathtt{B})$ car $\mathtt{P}(\mathtt{B}) = \mathtt{P}(\mathtt{B} \sqcup \mathtt{H}) + \mathtt{P}(\mathtt{B} \sqcup \neg\mathtt{H})$. Ce qui classe les règles du cas (c) en premier.

La dépendance

La **dépendance** est utilisée pour mesurer une distance de la confiance de la règle par rapport au cas d'indépendance de $\mathtt{B}$ et $\mathtt{H}$.

$$\mathtt{dep}\ [\mathtt{B} \longrightarrow \mathtt{H}] = |\mathtt{P}(\mathtt{H}|\mathtt{B}) - \mathtt{P}(\mathtt{H})| \tag{2.3}$$

Cette mesure varie dans l'intervalle $[0, 1[$. Plus cette mesure est proche de 0 (resp. 1) plus $\mathtt{B}$ et $\mathtt{H}$ sont indépendants (resp. dépendants). Ce qui augmente le plus sa valeur est la taille de $\mathcal{D}(\mathtt{H})$. Les valeurs sont sensiblement égales pour les cas (a) et (b). C'est particulièrement notable pour les règles exactes où la confiance $\mathtt{P}(\mathtt{H}|\mathtt{B})$ vaut 1 et donc $\mathtt{dep}\ [\mathtt{B} \longrightarrow \mathtt{H}] = 1 - \mathtt{P}(\mathtt{H})$ ne dépend pas de $\mathtt{P}(\mathtt{B})$. Par conséquent, la dépendance permet de séparer les règles du cas (c) d'une part et des cas (a) et (b) d'autre part.

La nouveauté et la satisfaction

La **nouveauté** est définie par :

$$\mathtt{nov}\ [\mathtt{B} \longrightarrow \mathtt{H}] = \mathtt{P}(\mathtt{B} \sqcup \mathtt{H}) - \mathtt{P}(\mathtt{B}) \times \mathtt{P}(\mathtt{H}) \tag{2.4}$$

La valeur absolue de cette mesure vaut $\mathtt{dep}\ [\mathtt{B} \longrightarrow \mathtt{H}] \times \mathtt{P}(\mathtt{B})$. Plus $\mathtt{P}(\mathtt{B})$ est faible, plus cette mesure est faible. Ainsi, les règles des cas (b) sont en fin de classement et sont différenciées du cas (a), alors que la dépendance ne le fait pas.

La nouveauté varie entre $]-1, 1[$ et prend une valeur négative quand $\mathtt{P}(\mathtt{B} \sqcup \mathtt{H}) < \mathtt{P}(\mathtt{B}) \times \mathtt{P}(\mathtt{H})$. La nouveauté s'approche de -1 pour des règles de faibles supports $\mathtt{P}(\mathtt{B} \sqcup \mathtt{H}) \simeq 0$. Nous sommes intéressés [CNT09] par les petites valeurs absolues de cette mesure, autour de la valeur d'indépendance 0). La nouveauté est symétrique alors que la règle $\mathtt{B} \longrightarrow \mathtt{H}$ peut avoir plus de contre-exemples que la règle $\mathtt{B} \longrightarrow \mathtt{H}$. Pour cette raison, nous introduisons la **satisfaction** :

$$\mathtt{sat}\ [\mathtt{B} \longrightarrow \mathtt{H}] = \frac{(\mathtt{P}(\neg\mathtt{H}) - \mathtt{P}(\neg\mathtt{H}|\mathtt{B}))}{\mathtt{P}(\neg\mathtt{H})} \tag{2.5}$$

qui s'écrit également :
$|\mathtt{sat}\ [\mathtt{B} \longrightarrow \mathtt{H}]| = \frac{\mathtt{P}(\mathtt{H}|\mathtt{B}) - \mathtt{P}(\mathtt{H})}{1 - \mathtt{P}(\mathtt{H})} = \frac{\mathtt{dep}\ [\mathtt{B} \longrightarrow \mathtt{H}]}{\mathtt{P}(\neg\mathtt{H})}$ car $\mathtt{P}(\neg\mathtt{H}) - \mathtt{P}(\neg\mathtt{H}|\mathtt{B}) = (1 - \mathtt{P}(\mathtt{H})) - (1 - \mathtt{P}(\mathtt{H}|\mathtt{B}))$
$= \mathtt{P}(\mathtt{H}|\mathtt{B}) - \mathtt{P}(\mathtt{H})$, avec $\mathtt{P}(\mathtt{H}|\mathtt{B}) + \mathtt{P}(\mathtt{H}|\neg\mathtt{B}) = 1$.

Cette mesure varie dans l'intervalle $]-\infty, 1]$ et vaut 0 en cas d'indépendance de B et H. La satisfaction n'est pas utile pour classer les règles exactes car sa valeur est 1 (puisque les règles exactes ont une confiance $P(H|B) = 1$). Pour cette mesure, $P(H)$ apparaît au numérateur et au dénominateur, donc la variation de cette mesure dépend de $P(B)$. Plus $P(B)$ est faible, plus cette mesure est élevée. Par l'intermédiaire de cette mesure, les règles du cas (a) sont en fin de classement et sont différenciées du cas (b). Nous sommes intéressés par les valeurs élevées de cette mesure (autour de la valeur 1).

En résumé, ces deux mesures peuvent être consultées simultanément lorsqu'on se trouve dans les cas (a) ou (b) (pour des règles à faible dépendance). Plus la nouveauté est faible et la satisfaction forte, plus la règle est considérée comme significative.

2.4.2 Mesures dites "subjectives"

Sahar et Liu sont deux auteurs ayant proposé une approche subjective de l'évaluation des règles d'association. Je reprends de la thèse d'Hacène Cherfi que j'ai encadrée les principales caractéristiques de ces approches.

Dans [Sah99], l'analyste est mis à contribution pour classer les règles d'association selon quatre points de vue. Notons $B \longrightarrow H$ une règle donnée :

- Les règles vraies et non intéressantes : ce sont des règles dont les parties B et H ont une signification triviale. Par exemple, la règle époux $\longrightarrow$ marié. Lorsqu'une règle de ce type est rencontrée par l'analyste, il faut la garder et ne pas générer la famille de règles impliquant B' et H' tels que B' est un sur-motif de B (par rapport à l'inclusion $B \subset B'$) et H' est un sur-motif de H ($H \subset H'$) ;
- Les règles fausses et intéressantes : ce sont des règles dont la signification est fausse (homme $\longrightarrow$ marié), mais dont des sous-règles pourraient s'avérer intéressantes. Une sous-règle d'une règle est définie selon l'ordre d'inclusion des deux motifs : $B \sqcup H \subsetneq B' \sqcup H'$. Par exemple, les règles du type homme, possède un véhicule 4x4 $\longrightarrow$ marié sont présentées à l'analyste.
- Les règles fausses et non intéressantes : ce sont des règles dont la signification est fausse Salaire élevé $\longrightarrow$ marié mais dont la connaissance lorsqu'elle est augmentée d'autres termes en parties B et H n'intéresse pas l'analyste ;
- Les règles vraies et intéressantes : ce sont les règles idéales du point de vue de l'analyste. Toutes les sur-règles $B' \longrightarrow H'$ telles que $B' \sqcap H' \subsetneq B \sqcap H$ de cette règle sont présentées.

Un algorithme décide des règles candidates, celles qui seront présentées à l'analyste. L'analyste classe alors cette règle dans une des quatre catégories et la famille de règles proches est validée ou rejetée automatiquement.

La critique que nous formulons, à propos de cette méthode, est la perte potentielle de surrègles intéressantes sans prendre de précaution particulière. [LHC97, LHM99] ont proposé des heuristiques différentes. Des règles "de référence" sont classées selon 4 critères :

- Les règles d'association conformes aux connaissances du domaine ;
- Les règles d'association dont la partie B est conforme et la partie H non conforme, c'est-à-dire, qui sont surprenantes ;
- Les règles d'association dont la partie B est surprenante et la partie H conforme ;
- Les règles d'association dont les deux parties B et H sont surprenantes

Les autres règles sont alors comparées aux règles dites “general impressions” pour rapprocher une règle à analyser d'une règle de référence.

2.5 Application aux textes

De très nombreux travaux ont appliqué les règles d'association aux textes. Je n'en cite que quelques uns, ceux qui me semblent le plus en accord avec le projet de recherche que je présente au chapitre 5.

2.5.1 Classification de textes à l'aide de règles d'association

La classification de textes est un domaine où les méthodes numériques (Rocchio, K-Means ou SVN) donnent de très bons résultats alors que les méthodes symboliques donnent souvent des résultats un peu inférieurs. Cependant, les méthodes symboliques sont intéressantes lorsque les corpus d'apprentissage sont petits et elles donnent des résultats plus facilement interprétables par les experts qui peuvent éventuellement les enrichir. [VMR+07, LR09, LR10] appliquent aussi l'extraction de règles d'association à des textes ou segments de textes.

[VMR+07] s'inscrit dans la version 2007 du Défi sur la Fouille de Textes (DEFT 2007) et s'intéresse à la classification supervisée de fragments de textes en fonction des opinions exprimées dans ces textes. À ce titre, ces travaux s'inscrivent dans la lignée des travaux de Liu [LHM98] ou de [LHP01] où, dans la phase d'apprentissage, les règles d'association sont extraites des textes en fonction des différentes classes. Un nouveau texte est alors classé en fonction d'une heuristique exploitant une seule ou une combinaison de plusieurs de ces règles pour classer ce nouveau texte. [VMR+07] extrait des règles disjonctives pondérées par un χ^2. Cette pondération est alors utilisée pour calculer le score d'un nouveau texte par rapport à chacune des catégories.

[LR09, LR10] ont appliqué des méthodes similaires pour identifier des segments d'obsolescence dans des textes encyclopédiques, c'est-à-dire, des segments textuels qui contiennent de l'information susceptible d'évoluer dans le temps. L'article donne l'exemple de la règle suivante :

$premierParag.position : debutDivision \wedge zone.rubriqueName : NULL$
$\wedge title.entiteNom.classe : geopolitique \longrightarrow classe : obsol$

qui peut s'interpréter par la glose suivante : « une phrase qui contient une entité nommée de type géopolitique, qui est située dans un paragraphe en début de section et dont la rubrique textuelle est de type null (c.-à-d. rubrique non référencée) est classée comme obsolescente par le classifieur.»

2.5.2 Extraction de connaissances par règles d'association

Du point de vue de l'extraction de connaissances à partir de textes, Feldman, Dagan et Hirsh [FD95a, FH96] figurent parmi les premiers auteurs à appliquer des règles d'association directement au niveau du texte : les objets sont des textes et les propriétés sont les mots ou, par la suite, les termes que contiennent les textes [FFK+98]. Toutes les règles

d'association sont extraites par un outil nommé *FACT* à la façon de l'algorithme *A Priori* et [FH97] propose de naviguer dans l'ensemble des règles extraites à l'aide d'un langage permettant de définir des patrons, de façons similaire à klemettinen94. Une partie de ces travaux ont été repris dans l'ouvrage récent [FS07] mais l'ouvrage couvre la fouille de textes dans une acceptation qui dépasse la nôtre ou celle de [Hea04] : il n'ya pas réellement de distinction entre la recherche d'information et la construction de connaissances.

[FH96] a ainsi été expérimenté sur des dépêches Reuters. Les noms d'entreprise ou de sociétés ont préalablement été annotés ainsi que des mots ou verbes reflétant une vente, une cession, une fusion ou un achat.... Les règles d'association sont alors extraites, en utilisant des seuils de confiance et de support assez faibles. La confiance était alors utilisée pour qualifier la qualité d'une association de deux entreprises pour une opération financière. Si ces travaux ont le mérite de créer ce lien entre textes et règles d'association, il me semble qu'une méthode de type extraction d'information associée à la construction d'une base de données donnerait des résultats assez peu différents.

La thèse de Jérôme Azé [AR03, Azé03] porte également sur la fouille de textes par extraction de règles d'association. Le premier point qui retient mon attention dans ce travail porte sur la discrétisation des données. Contrairement à ce qui est quasi-consensuel en recherche d'information, il n'y a pas de consensus sur comment représenter le contenu d'un texte. Nous reviendrons sur ce point section 4.3.1. Le second point longuement développé dans cette thèse porte sur les mesures objectives associées aux règles d'association. Les propriétés essentielles de ces mesures sont bien exposées et finalement deux d'entre elles sont plus particulièrement expérimentées : la moindre contradiction et l'implication normalisée. Cependant, ces travaux restent très fortement ancrés dans l'utilisation de mesures statistiques pour classer et filtrer les règles d'association. La connaissance, telle que nous la définissons au début de ce mémoire a finalement peu de place dans tout le processus. C'est, à mes yeux, une des raisons pour lesquelles les résultats de ce travail de fouille de textes n'a eu auprès des experts des domaines d'expérimentation qu'un succès assez mitigé.

Les travaux sur la fouille de textes menés par l'AIFB en Allemagne par Alexander Maedche s'accordent beaucoup mieux avec notre approche [MS00b, MS00a, MS01b, MS01a, MS03]. Les règles d'associations sont extraites dans le but d'identifier des relations hiérarchiques (lien de type EST-UN) entre les concepts ou des relations transversales pour alimenter la construction d'une ontologie.

2.5.3 Extraction d'information par règles d'association

Quelques éléments sur les motifs séquentiels

Ce n'est que récemment, que l'on observe l'utilisation des règles d'association pour l'extraction d'information (EI), c'est-à-dire, pour l'apprentissage à partir des textes de patrons pour l'identification dans les textes de l'information "utile". Une des raisons vient probablement du fait que, nous le verrons en section 4.3, les méthodes utilisées en EI sont majoritairement numériques. La seconde raison est liée au développement relativement récents d'algorithmes efficaces pour traiter des données séquentielles. Il est alors possible d'extraire des phrases des motifs séquentiels, prenant en compte l'ordre des mots dans les

TABLE 2.2 – Exemple de données (transactions) séquentielles

Client	Date	item
C_1	01/01/2004	B,F
C_1	02/02/2004	B
C_1	04/02/2004	C
C_1	18/02/2004	H,I
C_2	11/01/2004	A
C_2	12/01/2004	C
C_2	29/01/2004	D,F,G
C_3	05/01/2004	C,E,G
C_3	12/02/2004	A,B
C_4	06/02/2004	B,C
C_4	07/02/2004	D,G
C_4	08/02/2004	I

phrases, ordre qui est essentiel pour l'identification, par exemple, de rôles sémantiques.

L'extraction de motifs séquentiels [Mas02, MTP04] a tout d'abord été posée dans le contexte des bases de données. Les bases de la recherche de motifs séquentiels présentés ci-dessous sont empruntés à [MTP04]. D'apparence très proche du problème de l'extraction de motifs, l'extraction de motifs séquentiels pose cependant des problèmes de complexité pouvant engendrer des temps de calcul prohibitifs.

Définition 5 (Transaction) *Une transaction constitue, pour un client C, l'ensemble des items achetés par C à une même date. Dans une base de données client, une transaction s'écrit sous la forme d'un ensemble : id-client, id-date, itemset. Un itemset est un ensemble d'items non vide noté $(i_1 i_2 ... i_k)$ où i_j , avec j de 1 à k, est un item (il s'agit de la représentation d'une transaction non datée). Une séquence est une liste ordonnée, non vide, d'itemsets notée $< s_1 s_2 ... s_n >$ où s_j est un itemset (une séquence est donc une suite de transactions qui apporte une relation d'ordre entre les transactions). Une séquence de données est une séquence représentant les achats d'un client. Soit $T_1, T_2, ..., T_n$ les transactions d'un client, ordonnées par dates d'achat croissantes et soit itemset(T_i) l'ensemble des items correspondants à T_i, alors la séquence de données de ce client est $< itemset(T_1) itemset(T_2) ... itemset(T_n) >$.*

Le client C_4 est associé à la séquence $< (BC)(DG)(I) >$ qui s'interprète comme *le client C_4 a acheté ensemble les items B et C, puis il a acheté simultanément les items D et G puis, encore plus tard, il a acheté l'item I.*

Définition 6 (Inclusion de séquences) *Soit $s_1 =< a_1 a_2 ... a_n >$ et $s_2 =< b_1 b_2 ... b_m >$ deux séquences de données. s_1 est incluse dans s_2 ($s_1 \prec s_2$) si et seulement si il existe $i_1 < i_2 < .... < i_n$ des entiers tels que $a_1 \subseteq b_{i_1}, a_2 \subseteq b_{i_2}, ..., a_n \subseteq b_{i_n}$*

La séquence $s_1 =< (C)(DE)(H) >$ est incluse dans la séquence $s_2 =< (G)(CH)(I)(DEF)(H) >$ car $(C) \subseteq (CH)$, $(DE) \subseteq (DEF)$, et $(H) \subseteq (H)$. En revanche $< (C)(E) > \not< (CE) >$.

De façon assez similaire au support pour les motifs, le support d'une séquence s est le nombre de clients dont la séquence inclut s. On retrouve des propriétés proches de celles sur les motifs. Soient deux séquences s_1 et s_2 telles que $s_1 \prec s_2$, alors le support de s_1 est supérieur ou égal à celui de s_2. De même, toute séquence incluant une séquence non fréquente est non fréquente.

L'algorithme GSP est un des premiers algorithmes développés suivant des principes proches de ceux de l'algorithme d'extraction de motifs A PRIORI. Nous ne détaillerons pas ici les optimisations qui ont été apportées par la suite qui sont clairement exposés dans [MTP04].

Apprentissage de patrons pour l'extraction d'information

L'extraction d'information à partir de textes représente un enjeu majeur pour la fouille de textes. C'est une tâche malheureusement très longue et coûteuse voire même fastidieuse qui consiste à repérer les fragments de textes contenant une information "utile" (qui est bien sûr dépendante de la tâche à réaliser), à identifier des patrons caractérisant ces fragments puis à appliquer ces patrons pour une recherche exhaustive dans les textes. La plupart des approches d'apprentissage reposent sur des méthodes numériques qui ont un fonctionnement de type "boite noire". D'autres approches [PCZ+02] reposent sur des outils du traitement automatique des langues, notamment des analyseurs syntaxiques, mais, malgré des outils de plus en plus performants, elles se heurtent encore à la complexité des énoncés.

L'équipe du GREYC à Caen a mené une série de travaux et d'expérimentations très prometteurs dont, par exemple, [PC09, PCK+09]. Ces travaux proposent d'identifier les entités nommées et les interactions entre gènes par extraction de motifs/règles séquentiels.

Dans [PCK+09], une première étape consiste à identifier les noms de gènes dans les phrases et à les "anonymiser" en replaçant ces noms par un même nom générique "AGENE". Puis un certain nombre de contraintes sont prises en compte :

- un motif doit inclure au moins deux noms de gènes,
- un motif doit contenir un verbe,
- seules les séquences fréquentes maximales sont extraites pour réduire la redondance.

Malgré ces contraintes, il s'avère encore nécessaire de réduire le nombre de motifs séquentiels. Pour chaque nom ou verbe impliqué dans une séquence, l'objectif va être de ne garder que k (fixé à 4 dans l'expérimentation) motifs. La réduction du nombre de motifs se fait alors par l'extraction récursive proposée par [Sou07].

Les résultats de ces travaux donne une bonne précision et un rappel moyen, résultats que les auteurs considèrent comme comparable aux autres travaux. Les auteurs envisagent des améliorations assez naturelles à leur méthode, notamment la prise en compte de contraintes et de données linguistiques plus fortes, notamment, la modalité puisque l'on sait que la négation est un facteur de bruit important pour de tels outils.

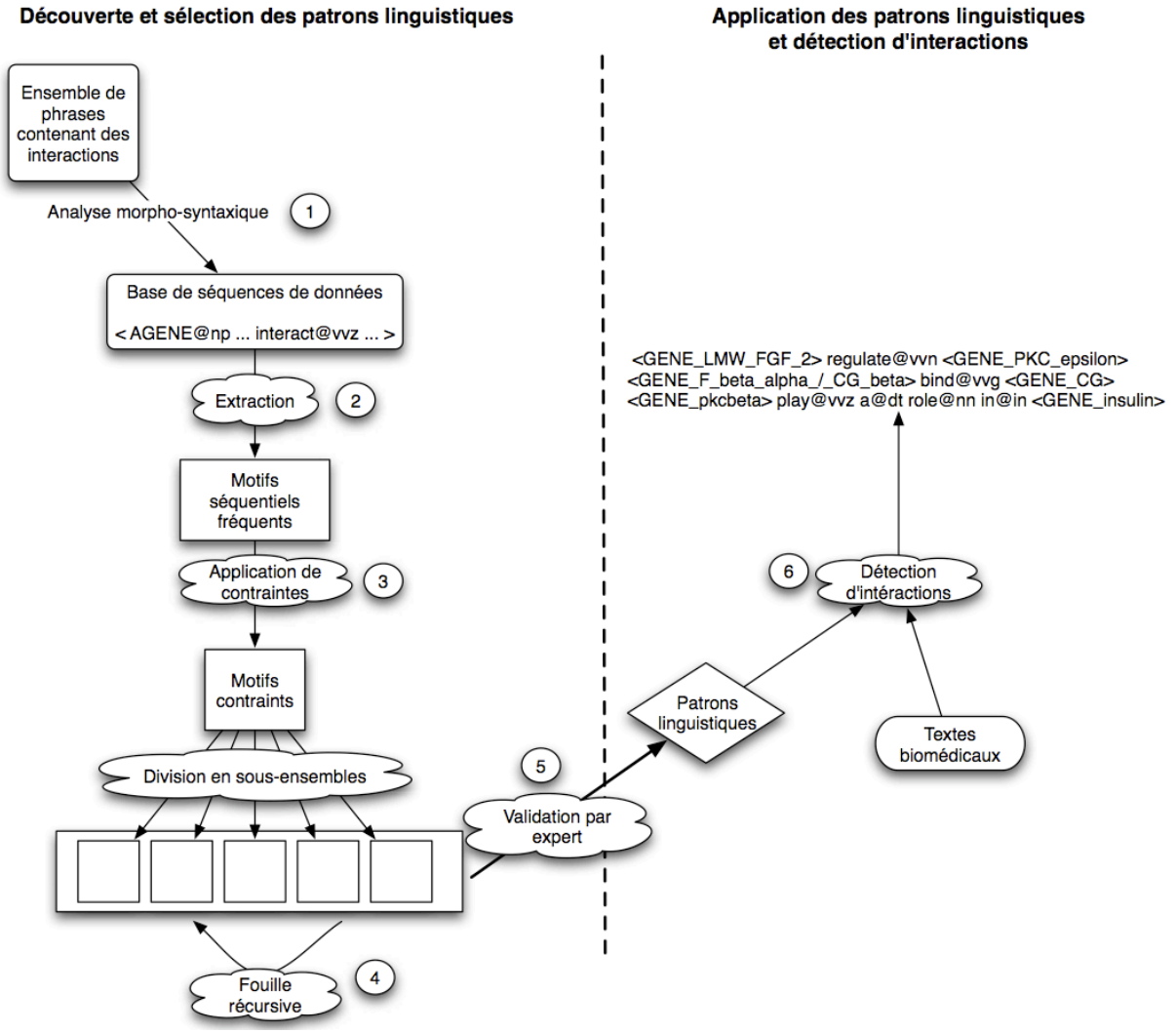


FIGURE 2.2 – Schéma général pour l'apprentissage de patrons séquentiels pour l'extraction d'information [PCK+09].

2.6 Ma contribution sur les règles d'association

2.6.1 Motifs rares et règles rares

La plateforme de fouille de données CORON développée par Laszlo Szathmary pendant sa thèse ([Sza06]) avec Amedeo Napoli, intègre un algorithme d'extraction de motifs rares. J'ai participé à la réflexion sur l'intérêt des motifs rares dans la fouille de données et à la définition d'algorithmes d'extraction de ces motifs rares [2].

Les motifs fréquents sont la cible de nombreux travaux, tant sur les problèmes algorithmiques que sur leur application dans différents domaines. L'idée développée ici est que les motifs rares puissent également être intéressants [SMP+06, MNSY06]. Considé-

2. CORON (http ://coron.loria.fr) est un logiciel déposé dont les auteurs sont L. Szathmary (60%) et A. Napoli (20%) et Y. Toussaint (20%)

rons l'exemple factice d'une base de données médicales et le problème de l'identification de la cause de maladies cardio-vasculaires (MCV). Une règle d'association fréquente (extraite d'un motif fréquent) comme `niveau élevé de cholestérol` $\longrightarrow$ `MCV` permet de faire émerger l'hypothèse que les individus ayant un fort taux de cholestérol ont un risque élevé de MCV. À l'opposé, s'il existe un nombre conséquent de végétariens dans la base de données, alors une règle d'association rare comme `vegetarien` $\longrightarrow$ `MCV` permet de faire émerger l'hypothèse qu'un végétarien a un risque faible de contracter une MCV. Dans un tel cas, les motifs `végétarien` et `MCV` sont tous deux fréquents, mais le motif `végétarien, MCV` est lui-même rare. Il est à noter que la règle `végétarien` $\longrightarrow$ `MCV` a alors un support faible et une confiance faible.

Dans le cas général, un motif est fréquent si son support est supérieur à un seuil `maxsupp` et il est rare si son support est inférieur à un seuil `minsupp`. Cependant, dans le cas de [SMP+06, MNSY06], nous avons étudié la situation où il n'existe qu'une seule frontière entre les fréquents et les non-fréquents. Les motifs qui sont extraits sont les motifs rares minimaux (MRM), c'est-à-dire les motifs dont tous les sous-motifs ne sont pas rares. L'ensemble des motifs rares mininaux forme un ensemble générateur minimal à partir duquel tous les motifs rares peuvent être retrouvés.

Les motifs rares sont nombreux et comptent notamment, tous les motifs de support zéro (les motifs zéros), ceux qui n'apparaissent jamais. Les motifs générateurs zéro-minimaux sont les motifs zéros pour lesquels tous ses sous-motifs sont des motifs non-zéros. Le fait qu'ajouter un élément à un sous-motif non-zéro donne un motif zéro est porteur d'une information intéressante.

La plupart des travaux sur les motifs rares ([KR05, TSB09, SNV07]) se sont focalisés sur des questions algorithmiques pour optimiser l'extraction des motifs rares. Dans sa thèse au sein de l'équipe Orpailleur, Sandy Maumus a analysé les motifs rares extraits d'une cohorte de personnes mais l'intérêt des motifs rares n'est pas encore réellement probant. Pourtant, dans le cadre de textes, on ne peut s'empêcher de faire le lien avec les signaux faibles ainsi dénommés dans la veille technologique. On pourrait en effet rechercher les associations rares dans les textes, associations qui refléteraient de nouveaux usages de termes déjà existants pour décrire de nouveaux phénomènes, avant que, finalement, un terme ne soit créé pour identifier cette nouvelle notion. Malheureusement, une heuristique aussi grossière a peu de chance d'aboutir. On observe souvent une forte dispersion au niveau des termes ou, plus généralement, des structures linguistiques et les signaux faibles risquent de se trouver noyés dans beaucoup de bruit.

2.6.2 Classification de règles d'association selon un modèle de connaissances

Les travaux que j'ai réalisés sur la fouille de textes par extraction de règles d'association ont été menés dans le cadre de la thèse d'Hacène Cherfi. L'objectif était de pouvoir extraire à partir de textes sur le domaine de la microbiologie des éléments de connaissances sur le phénomène de résistance des bactéries aux antibiotiques par mutation génétique. Ces travaux ont fait l'objet de plusieurs publications. **Pour éviter les redites, je ne donne ici que les lignes directrices de ces travaux. J'invite donc le lecteur à lire**

en premier lieu notre publication placée en annexe A avant de poursuivre la lecture de ce chapitre. Cet article est un chapitre de livre sur le post-traitement de l'extraction de règles d'association [CNT09] et généralise les travaux introduits à la conférence ECAI en 2004 (Valencia, Espagne). Je souligne dans les sous-sections suivantes trois points originaux que nous avons développés dans ces travaux.

Prise en compte de la variation terminologique

Le contenu des textes est représenté, comme dans les autres travaux cités sur les règles d'association, par un ensemble de termes. La relation d'incidence du contexte formel associe donc à un texte les termes qu'il contient. Pour réduire la dispersion liée à des formes linguistiques de termes différentes, nous avons utilisé FASTR [Jac94] qui prend en compte la **variation terminologique**. FASTR identifie des variations non triviales et linguistiquement motivées pour les ramener à leur terme préférentiel. Ainsi la phrase "l'alimentation hydrique de l'arbre varie" est reliée au terme "variation de l'alimentation hydrique de l'arbre". Ce prétraitement des textes réduit la dispersion des données en évitant la caractérisation des textes par des termes différents référant au même concept. Il renforce également le poids des termes (nombre d'occurrences) dans le corpus.

Combinaisons de mesures objectives

Une première approche pour mettre en valeur certaines règles d'association, consiste à combiner les propriétés des différentes mesures. Nous avons proposé un algorithme d'ordonnancement des règles d'associations combinant 4 mesures objectives : l'intérêt, la conviction, la dépendance et la satisfaction [CT02a, CT02e, CT02b, CT02c, CNT03a, CNT03b, CT03, CNT06].

Définition de la mesure de vraisemblance

Les mesures objectives associées aux règles d'association ne dépendent que de la distribution des termes dans les textes et ne sont donc pas reliées à un modèle de connaissance qui représenterait les connaissances de l'expert. Pour qu'une mesure prenne en compte les connaissances de l'expert, il faut introduire dans le processus un modèle de connaissances. Nous avons ainsi défini la vraisemblance pour mesurer à quel point une règle d'association véhicule une connaissance déjà exprimée dans le modèle de connaissance [CJNT04, CNT05a, CNT09]. L'expert peut ainsi étudier les règles pour lesquelles l'indice de vraisemblance est faible en premier lieu puisqu'elles se "démarquent" des connaissances exprimées dans le modèle. Le second intérêt d'une telle approche est son incrémentalité : lorsque l'expert introduit de nouvelles connaissances dans le modèle, l'ordonnancement des règles selon la vraisemblance change.

2.6.3 Hiérarchisation de règles d'association

Ce travail aborde deux problèmes. D'une part, il s'agit de faciliter, pour un expert, l'accès au très grand nombre de règles d'association extraites à partir d'un contexte formel, en structurant ces règles de façon hiérarchique. D'autre part, il s'agit d'extraire des règles

d'association à partir de propriétés qui sont elles-mêmes hiérarchisées. Dans ce dernier cas, il s'agit de privilégier parmi plusieurs règles celle qui est la plus 'générale". Ce travail a fait l'objet du DEA de Rokia Bendaoud que j'ai encadré en 2004. La démarche s'inspire des deux algorithmes Basic Cumule introduits dans [SA95] et des travaux de Maedche et Staab [MS00a].

Subsomption dans le cas de propriétés non hiérarchisées

Dans le cas de propriétés non hiérarchisées, nous avons créé des classes d'équivalence de règles reposant sur l'extension de la règle, i.e. pour une règle $\mathtt{A} \longrightarrow \mathtt{B}$ l'ensemble des individus vérifiant la condition $\mathtt{A} \cup \mathtt{B}$. Cela revient à associer chaque règle d'association à un concept du treillis de Galois construit à partir du même contexte formel et à exploiter la relation de subsomption définie dans les treillis pour construire des classes d'équivalence entre règles et ainsi, les ordonner [Ben04, BTN05]. Appliqué à la base "zoo" [For] utilisée pour tester des méthodes de fouille de données, les 4 règles suivantes appartiennent à la même classe d'équivalence :

- r21:4pattes $\longrightarrow$ denté, respire, vertébré
- r32:denté, respire, vertébré $\longrightarrow$ 4pattes
- r5:4pattes, vertébré $\longrightarrow$ denté, respire
- r6:4pattes, respire $\longrightarrow$ denté, vertébré

Subsomption dans le cas de propriétés hiérarchisées

Définition 7 *Soient deux règles* $\mathtt{r1}:\mathtt{A} \longrightarrow \mathtt{B}$ *et* $\mathtt{r2}:\mathtt{C} \longrightarrow \mathtt{D}$ r1 *subsume* r2*, noté* $\mathtt{r2} \sqsubseteq \mathtt{r1}$ *si et seulement si une des conditions suivantes est vérifiée :*

1. C *est ancêtre de* A *et* $\mathtt{B} = \mathtt{D}$ *:* $\mathtt{A} \longrightarrow \mathtt{B} \sqsubseteq \hat{\mathtt{A}} \longrightarrow \mathtt{B}$.
2. D *est ancêtre de* B *et* $\mathtt{A} = \mathtt{C}$ *:* $\mathtt{A} \longrightarrow \mathtt{B} \sqsubseteq \mathtt{A} \longrightarrow \hat{\mathtt{B}}$.
3. B *est ancêtre de* A *et* D *est ancêtre de* B *:* $\mathtt{A} \longrightarrow \mathtt{B} \sqsubseteq \hat{\mathtt{A}} \longrightarrow \hat{\mathtt{B}}$

Lors d'une généralisation de la partie droite d'une règle, il n'est pas nécessaire de recalculer le support et la confiance. La règle ainsi généralisée sera nécessairement valide. En revanche, la généralisation de la partie gauche suppose que l'on vérifie que la confiance de la règle est supérieure au seuil σ_c pour s'assurer que cette règle est valide.

2.7 Conclusion

L'extraction de motifs ou l'extraction de règles d'association sont des méthodes très intéressantes en fouille de données qui trouvent dans le domaine du texte, des applications aux résultats prometteurs. Si les travaux initiaux sur les textes ([FD95b]) ne prenaient pas en compte les spécificités de la langue et traitaient les textes comme des sacs de mots, les travaux que nous avons cités sur la classification de textes ou encore sur l'extraction d'information montrent que l'on peut envisager de se confronter à une réalité linguistique plus "affutée". Ils permettent d'apprendre des connaissances linguistiques qui peuvent ensuite être exploitées à la fois par des outils automatiques, des experts du domaine ou

des linguistes. Ils présentent l'avantage de pouvoir être interprétés, modifiés et enrichis. Il s'agit donc là d'une alternative ou d'une complémentarité intéressante aux méthodes numériques utilisées pour ces mêmes tâches (voir section 4.3).

Du point de vue de l'extraction de connaissances, la lecture des règles par un expert du domaine et leur interprétation est assez aisée : l'ensemble des objets ayant permis l'extraction de cette règle est facilement identifiable et une règle met généralement en jeu un nombre assez limité d'attributs. Chaque règle peut être lue et interprétée indépendamment des autres, même si, comme nous l'avons évoqué, il est aussi possible de classer les règles les unes par rapport aux autres.

Il subsiste deux grandes difficultés. La première est le très grand nombre de règles extraites. Des solutions ont été envisagées : réduction du nombre de règles (restriction aux motifs fermés, aux règles informatives, aux bases de règles...), tri des règles selon des indices objectifs ou subjectifs, utilisation de patrons de règles.

La seconde difficulté est liée à l'interprétation même des règles. D'une part, une règle d'association n'est pas une relation entre une cause et une conséquence. En pharmacovigilance, par exemple, on recherche les médicaments qui induisent des effets indésirables (ce qui est appelé un signal). Ainsi une règle `médicament` $\longrightarrow$ `effet-indésirable` signifie littéralement "Les patients qui ont pris `médicament` ont perçu `effet-indésirable` avec un support `s` et une confiance `c`". Mais une règle `effet-indésirable` $\longrightarrow$ `médicament` avec une confiance `c` signifie que "parmi les personnes qui ont perçu `effet-indésirable` `c` pour cent avaient pris `médicament`. Cette règle peut donc aussi traduire un signal en pharmacovigilance.

Appliquer des filtres sur un ensemble de règles est une méthode assez intuitive pour réduire le nombre de règles. Cependant, appliquer des filtres sur une base de règles au lieu de les appliquer sur l'ensemble des règles informatives, par exemple, peut entraîner du silence : certaines règles, pourtant valides, ne seront pas extraites. On pourra alors préférer une extraction de motifs ou de règles sous contraintes comme le propose par exemple [LFF09].

L"extraction de règles d'association est un outil très intéressant pour la fouille de données avec déjà beaucoup d'applications aux textes. Elles ont notamment été utilisées pour construire une hiérarchie de concepts comme dans [CHS05]. En revanche, les règles d'association semblent assez mal appropriées à la construction d'une base de connaissances, base dans laquelle les concepts seraient définis par des propriétés et des relations entre eux. L'analyse formelle de concepts que nous présentons dans la section suivante, nous semble, de ce point de vue, beaucoup plus appropriée.

Chapitre 3

L'analyse formelle de concepts

L'analyse formelle de concepts [GW99] est une méthode de classification permettant de construire des concepts et des hiérarchies de concepts à partir d'un ensemble d'individus décrits par un ensemble d'attributs. Ces travaux reposent à la base sur la théorie des treillis [Bir97, BM70]. La communauté FCA est très active tant sur le plan théorique que sur l'application pratique de ces travaux. Nous ne reprendrons dans ce document que quelques définitions essentielles à la compréhension de notre travail et de nos perspectives.

3.1 Quelques éléments de base en FCA

Définition 8 (Contexte formel, rappel) *Un* **contexte formel** *est un triplet* $\mathbb{K} = (\mathtt{O}, \mathtt{A}, \mathtt{I})$*, où* $\mathtt{O}$ *est un ensemble d'objets,* $\mathtt{A}$ *est un ensemble d'attributs et* $\mathtt{I} \subseteq \mathtt{O} \times \mathtt{A}$ *est la relation d'incidence.* $\mathtt{oIa}$ *signifie que l'objet* $\mathtt{o}$ *possède l'attribut* $\mathtt{a}$*.*

Deux opérateurs de dérivation notés $'$ relient les ensembles d'objet et d'attributs [GW99]. Soient $\mathtt{X} \subseteq \mathtt{O}$ et $\mathtt{Y} \subseteq \mathtt{A}$: $\mathtt{X}' = \{\mathtt{a} \in \mathtt{A} | \forall \mathtt{o} \in \mathtt{X}, \mathtt{oIa}\}$, $\mathtt{Y}' = \{\mathtt{o} \in \mathtt{O} | \forall \mathtt{a} \in \mathtt{Y}, \mathtt{oIa}\}$ Les deux composées de ces opérateurs, notées $''$, sont des opérateurs de fermeture sur $2^{\mathtt{O}}$ et $2^{\mathtt{A}}$, c'est-à-dire que, $\forall \mathtt{A}, \mathtt{B} \in 2^{\mathtt{O}} (resp. 2^{\mathtt{A}})$,

- (i) $\mathtt{A} \subseteq \mathtt{A}''$ (extensivité),
- (ii) $\mathtt{A} \subseteq \mathtt{B}$ then $\mathtt{A}'' \subseteq \mathtt{B}''$ (monotonicité),
- (iii) $(\mathtt{A}'')'' = \mathtt{A}''$ (idempotence),
- $\mathtt{A}$ est dit *fermé* si $\mathtt{A}'' = \mathtt{A}$.

Définition 9 *Un* **concept formel** *est une paire* $(\mathtt{X}'', \mathtt{X}')$ *où* $\mathtt{X} \subseteq \mathtt{O}$ *où* $\mathtt{X}''$ *est appelé extension du concept et* $\mathtt{X}'$ *l'intension.*

L'ensemble des concepts $\mathcal{C}_{\mathbb{K}}$ du contexte $\mathbb{K}$ est partiellement ordonné selon l'inclusion de l'extension : $(\mathtt{X}_1, \mathtt{Y}_1) \leq_{\mathbb{K}} (\mathtt{X}_2, \mathtt{Y}_2) (\Leftrightarrow \mathtt{X}_1 \subseteq \mathtt{X}_2)$.

La structure $\mathcal{L} = \langle \mathcal{C}_{\mathbb{K}}, \leq_{\mathbb{K}} \rangle$ est un treillis de concepts ou treillis de Galois.

La table 3.1 reprend le contexte formel de la section précédente. Le treillis de Galois, construit à partir de ce contexte est donné par la figure 3.1. L'intension d'un concept est donné par l'ensemble $\mathtt{I}$, son extension par l'ensemble $\mathtt{E}$. L'ensemble $\mathtt{G}$ est l'ensemble des motifs générateurs de $\mathtt{I}$, c'est-à-dire, les motifs minimaux dont la fermeture est égale à $\mathtt{I}$.

TABLE 3.1 – Exemple d'un contexte formel où $\{\mathtt{a}, \mathtt{b}, \ldots\}$ sont des termes et $\{\mathtt{d}_1, \mathtt{d}_2, \ldots\}$ sont des textes.

$\mathcal{I}$	a	b	c	d	e
$\mathtt{d}_1$	0	1	1	0	1
$\mathtt{d}_2$	1	0	1	1	0
$\mathtt{d}_3$	1	1	1	1	0
$\mathtt{d}_4$	1	0	0	1	0
$\mathtt{d}_5$	1	1	1	1	0
$\mathtt{d}_6$	1	0	1	1	0

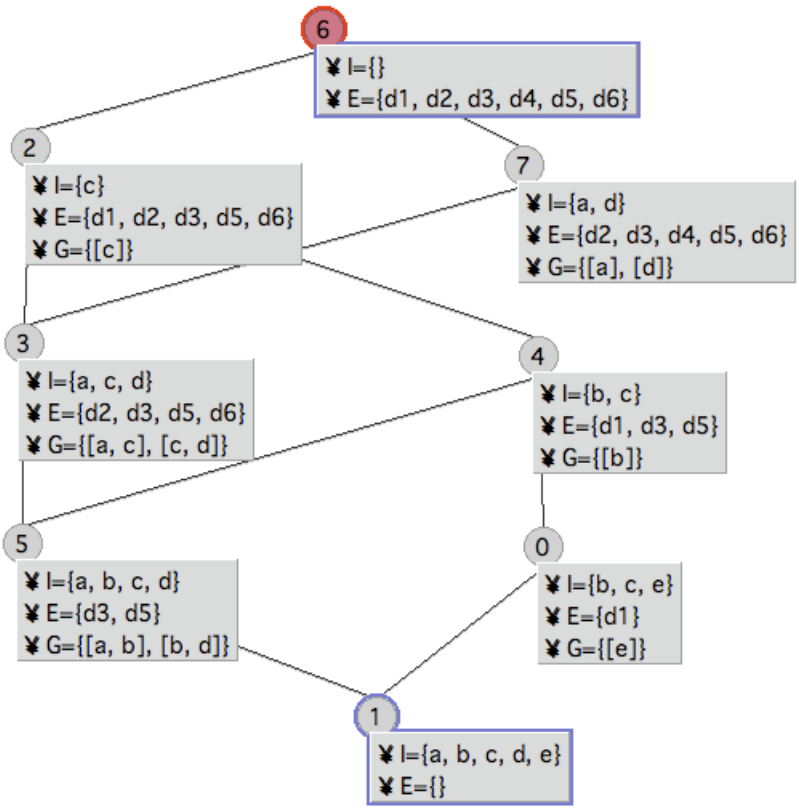


FIGURE 3.1 – Treillis de concepts associé au contexte donné par le contexte 3.1.

Nous empruntons à [RH08] quelques propriétés souvent citées à propos de la FCA car elles sont importantes dans le contexte de la construction d'ontologie :

- exhaustivité : le treillis contient tous les regroupements significatifs d'attributs et d'individus. Toutes les combinaisons potentielles ou bien l'espace de recherche (par exemple en recherche d'information) est donné par le treillis qui permet de trouver la solution à une requête ou une solution approchée en utilisant l'ordre partiel du treillis.
- maximalité : L'intension et l'extension des concepts sont construits de telle façon que l'extension soit l'ensemble maximal des individus qui partagent l'ensemble maximal de propriétés qu'est l'intension.
- factorisation maximale : au sein du treillis, les propriétés peuvent n'être localisées qu'une seule fois, dans leur concept maximal (dit concept-attribut) ; tout concept inférieur au concept-attribut possède nécessairement la propriété ce qui permet de l'omettre dans leur visualisation, à l'instar de l'héritage entre classes dans le paradigme objet. C'est une propriété qui nous sera très utile lorsque l'on s'intéressera à proposer des définitions pour les concepts d'une ontologie.

3.1.1 Les algorithmes de construction de treillis

On retrouve pour les treillis les mêmes difficultés que pour la recherche de motifs fréquents : le nombre de concepts dans un treillis devient vite très grand, tout en étant borné par $2^{\min(|\mathtt{O}|,|\mathtt{A}|)}$. Construire un treillis est donc coûteux en temps et en mémoire. Il existe de nombreux algorithmes pour la construction de treillis ; certains d'entre eux recherchent en premier lieu les motifs fermés (fréquents) ou/et les motifs générateurs minimaux puis calculent ensuite la relation d'ordre entre ces ensembles. On sépare traditionnellement ces algorithmes en deux familles : les algorithmes non incrémentaux [Gan84, Bor86] et les algorithmes incrémentaux [GM94, VRHM03]. Plus récemment, on peut également citer [SVNG08] qui est non incrémental.

3.2 Ma contribution sur la FCA pour la fouille de données

L'article en annexe B présente les travaux que j'ai mené en pharmacovigilance, avec Jean Villerd, post-doctorant que j'ai encadré dans le contexte du projet Vigitermes.

Les systèmes de déclaration spontanée en pharmacovigilance collectent les situations (les cas) dans lesquelles des médicaments créent des effets indésirables chez un patient. Ces bases sont alors analysées pour créer des signaux, associations médicament-effet indésirable qui satisfont des critères statistiques établis. Les algorithmes d'extraction de signaux actuellement utilisés pour fouiller ces bases de données reposent donc sur le calcul de mesures de disproportionnalité : le risque relatif (RR, Relative risk) ou encore le PRR (Proportional Reporting Ratio). Des algorithmes très performants [LFF09] reposent sur l'extraction de motifs avec contraintes pour maximiser le RR. Dans ces travaux, les motifs les plus courts sont recherchés et seuls les motifs pour lesquels tous ses sous-motifs

ont une valeur de RR inférieure, sont gardés. Seul, un sous-ensemble des motifs générateurs est ainsi extrait. De plus, les auteurs proposent de tester une condition d'arrêt qui assure que les sur-motifs d'un motif ne seront jamais des motifs à garder. Une généralisation de cette approche définie pour 1 médicament et 1 effet indésirable pourrait traiter les situations avec plusieurs médicaments. Cependant le RR ou le PRR atteignent aussi leurs limites : le nombre de signaux est très important, certains signaux ne sont pas détectés et, à l'inverse, certains associations ne devraient pas être proposées, il n'est pas facile de relier des signaux (1 médicament, 1 effet indésirable) avec des associations plus complexes (plusieurs médicaments, plusieurs effets indésirables) impliquant ce même médicament, ou encore d'étudier l'impact des facteurs démographiques.

La FCA apporte une réponse à ces différentes questions par la construction d'un treillis dans lequel les objets sont les patients, et les attributs sont les médicaments, les effets indésirables et les facteurs démographiques. Les mesures statistiques habituellement utilisées peuvent également être calculées. Les expérimentations que nous avons menées ont montré que l'amélioration du processus de fouille passe maintenant par l'introduction de connaissances expertes et l'intérêt de la FCA réside dans cette possiblité d'introduire des connaissances expertes pour améliorer l'identification des signaux et, plus généralement, des associations pertinentes.

3.3 Application de la FCA aux textes

L'idée de faire émerger de textes des concepts, c'est à dire des classes d'objets que l'on peut définir ou caractériser par un ensemble d'attributs suscite beaucoup d'intérêt. Le premier cadre applicatif est venu de la recherche d'information pour s'étendre par la suite à la construction d'ontologie.

3.3.1 Application à la recherche d'information

C. Carpineto et G. Romano [CR96] sont parmi les premiers à expérimenter la FCA pour la recherche d'information. L'idée est simple et correspond à une application directe de la FCA : les objets sont les textes d'une base documentaire et ils sont caractérisés par les mots-clés qu'ils contiennent. Le treillis définit l'espace de recherche qu'il est possible d'interroger. La requête est elle aussi formulée par un ensemble de mots-clés. Le calcul de la réponse à une requête consiste à classer la requête dans le treillis. S'il existe un concept dont l'intension correspond exactement à la requête, la réponse est donnée par l'extension du concept. Sinon, les concepts subsumants ou plus généralement, les concepts voisins du concept requête seront des réponses approximatives satisfaisantes.

Le pari était osé puisque la recherche d'information avait de longue date opté pour une représentation vectorielle des textes, c'est à dire une liste de mots-clés pondérés généralement par le TF-IDF, alors que l'approche proposée reposait sur une représentation booléenne (présence/absence d'un mot-clé). La comparaison avec une méthode numérique (BMR – Best Match Ranking) dans [CR00] montre que le classement proposé par une approche à base de treillis est très performant.

Plusieurs systèmes de recherche d'information se sont alors développés. Parmi eux,

REFINER [CR98] ne construit que les concepts voisins de la requête au lieu de construire le treillis entier, CREDO [CR04] construit un treillis à partir de documents extraits par Google sur le web. On peut également citer FooCA [Koe06] ou encore SearchSleuth [DP07] qui, un peu comme REFINER, met l'emphase sur la reformulation.

Enfin, CreChainDo [NT08c] est un outil développé par Emmanuel Nauer qui reprend l'idée de CREDO et exploite l'idée que les concepts du treillis correspondent à une vision synthétique d'un domaine et que le treillis peut alors être exploité pour guider de façon dynamique la recherche de documents sur le web. La démarche est itérative : un premier treillis est construit à partir d'un ensemble de textes extraits par une requête sur un moteur de recherche du web. L'utilisateur peut alors invalider des concepts qui ne s'inscrivent pas dans sa recherche (l'extension de ces concepts sont des documents "hors-sujet"), ou, au contraire, valider des concepts, permettant ainsi d'interroger le web avec une requête étendue, plus précise.

3.3.2 Construction d'ontologies à partir de textes

La construction d'ontologie que ce soit à partir de données, de textes ou d'interviews d'experts est une activité qui relève plutôt de l'ingénierie des connaissances mais qui est très liée à la linguistique (lorsqu'il s'agit d'extraire l'information des textes) ou encore, à l'informatique, lorsqu'il s'agit de représenter ces connaissances pour les partager et les rendre exploitables par des machines. Un certain nombre de méthodologies ont été proposées pour guider la construction des ontologies [SAA+99, NM01, GPFLC04, BLC96], chacune mettant l'accent sur des points qui leur sont propres comme le rôle de l'expert, la nature des ressources, l'enrichissement des ontologies ou leur modularité...

La FCA pour la construction d'ontologie

La FCA construit une conceptualisation d'un domaine à partir d'un contexte formel décrivant des individus par des attributs. Appliqué à la recherche d'information, c'est le regroupement de termes ou mots-clés pour décrire les textes qui permettaient de construire l'espace de recherche. En revanche, dans le cadre de la construction d'ontologies, il s'agit d'identifier dans les textes les objets du domaine et les propriétés qui leur sont associées. On se place donc dans une analyse fonctionnelle de la langue, proche des travaux de [Har68] exploitée également dans des travaux comme [FNR98]. La construction d'une ontologie par l'utilisation de la FCA est donc une démarche bottom-up, qui part des instances et de leurs descriptions pour en abstraire des classes et des propriétés associées aux classes. Une telle construction fait l'hypothèse d'un monde fermé, reposant sur une description extentionnelle du monde [Ser08]. Un certain nombre de problèmes y sont liés comme notamment la mise à jour ou l'enrichissement d'une ontologie dans le temps : lorsqu'une ontologie est partagée par une communauté, il n'est pas possible de supprimer l'un de ces concepts sans se préoccuper des conséquences que cela peut avoir sur, par exemple, des requêtes SPARQL impliquant ce concept.

Cimiano, Maedche et Staab [CSHT04, CHS05] proposent une application immédiate de la FCA pour le domaine du tourisme. Les objets génériques sont `les hôtels`, `les appartements`, `les voitures...` et les propriétés sont du type `réservable`, `pouvant`

`être loué`, `pouvant être conduit`.... La FCA construit alors une conceptualisation du domaine que les auteurs considèrent comme le haut de la hiérarchie de l'ontologie. Pour pallier le bruit lié aux étapes précédent la FCA comme l'extraction d'information et même le bruit inhérent aux textes (un mot utilisé à la place d'un autre, un mot qui n'est pas présent dans tel texte mais présent dans tel autre texte qui traite pourtant du même sujet...) une méthode de lissage est introduite en amont de la construction de treillis. Elle permet ainsi de réduire la dispersion des données. Une évaluation de l'ontologie résultante est également proposée, reprenant des méthodes proposées dans [MS01a]. Nous reviendrons sur la question de l'évaluation en lien avec nos travaux dans la section suivante et en section 5.5.1.

Enrichissement et fusion d'ontologies

R. Navigli et P. Velardi [NV06] enrichissent une ontologie de haut niveau en recherchant dans un glossaire les différentes instances de ses concepts et des relations qu'elles entretiennent entre elles. Le repérage des termes et leurs relations est faite par des patrons syntaxiques puis les données sont regroupées pour déterminer pour chaque relation le domaine et le co-domaine.

G. Stumme et A. Maedche [SM01b] se sont intéressés à la fusion de deux ontologies en s'appuyant sur des textes. Leur approche repose sur la FCA : deux contextes formels sont créés, un pour pour chacune des ontologies. Un contexte relie chaque texte aux concepts qu'il contient. La fusion des deux ontologies se fait grâce à l'apposition de contextes, opération que nous détaillons dans la section suivante.

3.4 Ma contribution à la construction d'ontologie à partir de textes

Les travaux que j'ai menés sur la construction d'ontologie à partir de textes s'inscrivent dans le cadre de la thèse de Rokia Bendaoud [Ben09]. **J'invite le lecteur à lire en premier lieu l'article que nous avons publié à la conférence Knowledge Engineering and Knowledge Management (EKAW) 2008 placé en annexe C**. Cet article montre que la FCA et la RCA constituent un cadre unifié pour la construction d'ontologie à partir de textes. Je reprends ci-dessous, mais de façon très synthétique les points forts de cet article et plus généralement les points forts de la méthode que nous proposons.

L'ontologie que je construis est une conceptualisation d'un domaine, construite à partir d'un ensemble de textes et d'autres formes de ressources que nous détaillons dans les sections suivantes. Cette ontologie est constituée d'un ensemble de concepts et de relations entre concepts, représentés dans un langage formel. Elle est destinée à être utilisée soit par un expert dont le but, en l'interrogeant par un langage de type SPARQL, est d'avoir accès à une vision synthétique de son domaine, soit par une machine qui peut alors raisonner sur les connaissances pour produire de nouvelles connaissances. En d'autres termes, l'ontologie que nous construisons est la transformation du résultat de la FCA ou de la RCA en une base de connaissances codée, par exemple, en logique de descriptions.

La construction d'une ontologie dépend à la fois du domaine et de la tâche à laquelle cette ontologie est destinée mais l'impact de ces deux dimensions sur le processus est assez peu exploré dans les travaux sur les ontologies. Dans notre cas, le domaine et la tâche sont pris en compte de deux façons : d'une part dans le choix des ressources qui seront exploitées pour la construction de l'ontologie, d'autre part, dans la construction du modèle qui servira de guide à la FCA et la RCA.

3.4.1 Le traitement de ressources hétérogènes

Construire une ontologie à partir de textes suppose, en premier lieu, de constituer un corpus. Au delà du choix du domaine et de son étendue (faut-il choisir le domaine de la microbiologie ou le sous-domaine spécifique de la résistance des bactéries aux antibiotiques ?), la nature des textes doit être définie : compte-rendus médicaux, fiches de rapport d'incident, articles ou résumés d'article scientifique...

En second lieu, il faut se poser la question de savoir si toutes les informations nécessaires à la construction d'ontologie sont présentes. Dans un article scientifique, on ne trouve pas l'information selon laquelle un quinolone est un antibiotique, cela fait partie de la connaissance implicite que le lecteur est supposé avoir. Dans un précis ou un manuel scolaire, en revanche, ces notions seront définies explicitement, parfois même avec des structures textuelles clairement identifiées (définitions, mise en gras du terme défini). Ainsi, dans le "Précis de géomorphologie [3]", un certain nombre d'énoncés définitoires peuvent être identifiés comme, par exemple, dans une phrase du type : *"un* **glacis** *est une accumulation de glace qui..."*. Dans ce précis, le terme défini est écrit en gras, et le marqueur *est un* introduit clairement la définition. Il faut ajouter à cela que le livre contient un index des définitions renvoyant pour chaque terme ou concept, la page où il est défini. Dans d'autres cas, il faut collecter des ressources complémentaires pour pallier les informations qui manquent dans les corpus initiaux. La FCA et la RCA permettent de traiter et d'intégrer ces différentes ressources que nous qualifions d'hétérogènes.

Les trois types de ressources que nous identifions sont les suivantes :

- des ressources de type thésaurus ou ontologiques, classant les objets du domaine dans des catégories : *une poire est un fruit...*
- des ressources caractérisant les objets par des propriétés : *Une poire est sucrée...*
- des ressources reliant des objets par des relations : *Bacilli résiste à la cefotaxime*

L'article placé en annexe C montre que la FCA et la RCA permettent de traiter ces différents types de ressources. Nous revenons dans la section suivante sur les principaux atouts de ces méthodes.

3.4.2 FCA et RCA, un cadre formel unifié pour la construction d'ontologies

L'article en annexe C présente le schéma global de la construction d'ontologie à partir de textes et notamment comment les contextes formels sont créés à partir des données. L'analyse formelle de concepts (FCA pour *Formal Concept Analysis*) et son extension,

3. Précis de géomorphologie, M. Derruau, Masson, 1988

l'analyse relationnelle de concepts (RCA pour *Relational Concept Analysis*), proposent alors un cadre formel particulièrement puissant pour intégrer les trois types de ressources évoquées. Un des points essentiels repose sur le fait que ces approches permettent de construire des concepts à partir de la description d'individus. Je rappelle ici quelques éléments importants sur lesquels reposent nos travaux.

La construction d'un modèle *a priori*

La FCA prend en entrée un contexte formel et la RCA, une famille de contextes relationnels. La FCA prend donc en entrée un ensemble unique d'individus. En revanche, la RCA peut potentiellement distinguer plusieurs ensembles d'individus qu'elle va structurer en plusieurs treillis reliés ensemble. Le modèle décide *a priori* de ce que sont les individus du domaine (un ou plusieurs ensembles), des attributs associés à ces individus et des relations entre ces individus. A l'heure actuelle, la définition du modèle est une tâche incontournable qui permet de construire un processus d'extraction d'informations des textes en amont du processus de conceptualisation. Le modèle détermine aussi la méthode de classification, et notamment, dans le cadre de la RCA, quelles sont les différentes familles d'individus à considérer (les antibiotiques, les bactéries...) et les relations que l'on veut définir entre ces objets. Le modèle est probablement un des principaux éléments de prise en compte de la tâche dans la construction de l'ontologie.

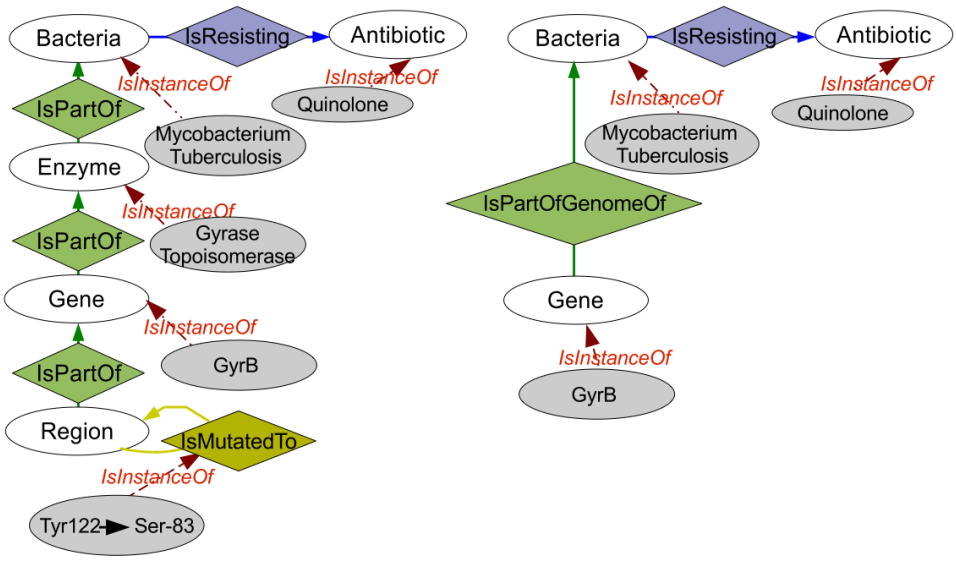


FIGURE 3.2 – A gauche, le schéma global décrivant les entités et relations en jeu dans le phénomène de résistance des bactéries aux antibiotiques, à droite, le schéma simplifié exploité pour nos expérimentations

L'apposition de contextes

Soient deux contextes $\mathbb{K}_1 = (\mathtt{G}, \mathtt{M}_1, \mathtt{I}_1)$ et $\mathbb{K}_2 = (\mathtt{G}, \mathtt{M}_2, \mathtt{I}_2)$, portant sur le même ensemble d'objets $\mathtt{G}$ et deux ensembles distincts d'attributs $\mathtt{M}_1$ et $\mathtt{M}_2$ ($\mathtt{M}_1 \cap \mathtt{M}_2 = \emptyset$). L'apposition en FCA permet de fusionner les deux contextes en un seul contexte [GW99].

Définition 10 *Soit* $\mathbb{K}_1 = (\mathtt{G}_1, \mathtt{M}_1, \mathtt{I}_1)$ *et* $\mathbb{K}_2 = (\mathtt{G}_2, \mathtt{M}_2, \mathtt{I}_2)$ *deux contextes formels. Si* $\mathtt{G} = \mathtt{G}_1 = \mathtt{G}_2$ *and* $\mathtt{M}_1 \cap \mathtt{M}_2 = \emptyset$*, alors* $\mathbb{K} = \mathbb{K}_1 | \mathbb{K}_2 = (\mathtt{G}, \mathtt{M}_1 \cup \mathtt{M}_2, \mathtt{I}_1 \cup \mathtt{I}_2)$ *est l'*apposition *des deux contextes* $\mathbb{K}_1$ *et* $\mathbb{K}_2$*.*

L'apposition permet donc de "fusionner" des sources d'information différentes, décrivant les mêmes objets mais par des propriétés différentes. Elle est utilisée notamment pour introduire des catégories pré-existantes, *i.e.* une classification consensuelle que les experts souhaitent prendre en compte ? Chaque objet est alors caractérisé par l'ensemble des catégories auxquelles il appartient, dans un contexte formel, et ces catégories sont alors intégrées à la construction *bottom-up* de l'ontologie.

La RCA

L'analyse relationnelle de concepts (Relational Concept Analysis, RCA) est présentée dans [HNRV07a] comme une extension de la FCA dans le but de prendre en compte dans la classification, les relations entre objets. Un concept formel est alors décrit par un ensemble d'attributs mais également par un ensemble de relations avec d'autres concepts.

Les données en RCA sont organisées en une famille de contextes relationnels (RCF) composée d'un ensemble de contextes $\mathbb{K}_i = (\mathtt{G_i}, \mathtt{M_i}, \mathtt{I_i})$ et d'un ensemble de relations $\mathtt{r_k} \subseteq \mathtt{G_i} \times \mathtt{G_j}$. Les ensembles $\mathtt{G_i}$ et $\mathtt{G_j}$ sont les ensembles d'objets des contextes $\mathbb{K}_i$ et $\mathbb{K}_j$, appelés respectivement *domaine* et *co-domaine* de la relation $\mathtt{r_k}$.

La RCA utilise le mécanisme du *relational scaling* pour définir des attributs relationnels. Pour une relation, $\mathtt{r} : \mathtt{G_i} \longrightarrow \mathtt{G_j}$, reliant des objets de $\mathtt{G_i}$ à des objets de $\mathtt{G_j}$, un attribut relationnel est créé, noté $\mathtt{r} : \mathtt{c}$, où $\mathtt{c}$ est un concept de $\mathcal{L}(\mathtt{G_j}, \mathtt{M_j}, \mathtt{I_j})$ – le treillis construit à partir du contexte $\mathbb{K}_j = (\mathtt{G_j}, \mathtt{M_j}, \mathtt{I_j})$. Alors, pour un objet $\mathtt{g} \in \mathtt{G_i}$, l'attribut relationnel $\mathtt{r} : \mathtt{c}$ caractérise la "corrélation" entre $\mathtt{g}$ et $\mathtt{r}(\mathtt{g}) = \mathtt{h}$ qui est une instance du concept $\mathtt{c} = (\mathtt{X}, \mathtt{Y})$ de $\mathcal{L}(\mathtt{G_j}, \mathtt{M_j}, \mathtt{I_j})$. Différents niveaux de corrélation peuvent être envisagés, une corrélation existentielle (ou existential scaling) où $\mathtt{r}(\mathtt{g}) \cap \mathtt{X} \neq \emptyset$, et une corrélation universelle où $\mathtt{r}(\mathtt{g}) \subseteq \mathtt{X}$.

Deux différentes heuristiques pour construire un treillis relationnel sont proposées : "Narrow encoding" et "wide encoding". Cette dernière, le "wide encoding" correspond probablement à l'heuristique la plus intuitive. Si, dans la classification finale, un attribut relationnel relie le concept $\mathtt{c_1}$ au concept $\mathtt{c_2}$ par une relation $\mathtt{r}$, cela signifie que chaque objet de l'extension de $\mathtt{c_1}$ est relié à au moins un objet de l'extension de $\mathtt{c_2}$ par la relation $\mathtt{r}$. Autrement dit, $\forall \mathtt{x} \in \mathtt{extend}(\mathtt{c_1}), \mathtt{r}(\mathtt{x}) \cap \mathtt{extend}(\mathtt{c_2}) \neq \emptyset$.

La représentation en logique de descriptions

Le treillis final résultant de la RCA peut être transformé en une base de connaissances codée en logique de descriptions comme le suggèrent [HNRV07a, RHNV$^+$08a, HNRHV07]. Cette transformation introduit des concepts primitifs et des concepts définis sur lesquels peut être appliqué un raisonneur pour résoudre, par exemple, une requête. La logique des descriptions cible permettant de représenter la famille de treillis construite par RCA est de type $\mathcal{FLE}$, qui contient les constructeurs $\top$ (top), $\bot$ (bottom), $\mathtt{C} \sqcap \mathtt{D}$ (conjonction de concepts), $\forall \mathtt{r}.\mathtt{C}$ et $\exists \mathtt{r}.\mathtt{C}$ (quantification universelle and existentielle).

3.4.3 Evaluation

L'évaluation de la qualité d'une ontologie est un problème difficile et qui reste, à mon sens, non résolu jusqu'à présent [GPFLC04, GW00b, VVSH08]. Sous l'angle méthodologique, cela reste aussi dans notre approche un point faible. Une des raisons est que nous percevons l'ontologie comme une vision synthétique mais partielle d'un domaine, pouvant d'ailleurs ne pas être consensuelle. Cela fait partie de mes perspectives de recherche que d'évaluer les ontologies relativement aux critères avancés dans les travaux cités ici (voir section 5.5.1). Dans le cadre de la thèse de Rokia Bendaoud, l'évaluation de l'ontologie repose sur les interactions que nous avons eues avec les experts en astronomie ou encore avec les experts en microbiologie. Dans certains cas, la comparaison des classes du treillis et de leur définition en intension peut aussi être un critère d'évaluation de la méthode [BTN08].

3.4.4 Interaction

L'interaction avec l'expert est un point essentiel dans la construction d'ontologie, point qui est d'ailleurs en partie lié à l'évaluation. C'est en sollicitant l'avis de l'expert sur les résultats d'une première passe qu'il est le plus à même de proposer des modifications de nature à améliorer l'ontologie pour la passe suivante.

Dans nos travaux, je conçois l'interaction comme la possibilité que l'on donne à l'expert de modifier les données qui sont fournies en entrées à la FCA, c'est à dire les contextes formels. Cette intervention a été manuelle jusqu'à présent, mais on peut envisager de mettre en oeuvre des modifications plus importantes de façon automatique. Les raisons pour lesquelles il n'est pas souhaitable que l'expert corrige l'ontologie résultante directement au niveau du code en logique de descriptions sont assez immédiates. La FCA et RCA assurent une parfaite cohérence de la base qui ne peut être garantie si l'on modifie à la main le résultat. Même si d'autres outils peuvent en vérifier la cohérence – comme par exemple, un raisonneur en logique de descriptions – une telle modification nous fera perdre la "traçabilité" entre les données de départ et l'ontologie résultante. L'interaction doit donc permettre de "customiser" les ressources pour les adapter aux besoins de l'expert.

Les raisons pour lesquelles l'expert peut souhaiter intervenir sur les données de départ sont multiples :

- l'expert n'est pas d'accord sur certains concepts de l'ontologie et la classification des objets pour une des raisons suivantes :
  - il y a du bruit dans les ressources et que ce bruit conduit à des regroupements indésirables ;
  - l'expert attend ou espère l'émergence de classes qu'il ne voit pas, simplement parce que les ressources (souvent collectées de façon opportunistes) sont peu appropriées à ses besoins ;
- certaines classes devraient être affinées, c'est-à-dire décomposées en sous-classes par la prise en compte de nouvelles propriétés ou relations, ou bien, au contraire, certaines propriétés qui ne s'avèrent finalement pas significatives devraient être soit regroupées avec d'autres soit supprimées.

Un expert peut par exemple identifier un problème, comme il l'a fait en microbiologie,

TABLE 3.2 – Opérations proposées à l'expert sur les différents types de ressources

Hierarchical description (thesaurus)	Attributes	Relations
Ajouter une nouvelle classe	Ajouter un attribut à un objet ou à tous les objets	Ajouter une relation entre des objets
Détruire une classe	Retirer un attribut à un objet ou à tous les objets	Détruire une relation entre objets

pour le concept `C205`. L'intension de ce concept est :
{`sticks`, `neutral gram`, `heterotrophic`, `immobile`, `Aerobic`, `Resist.A0`, ...}
et son extension : {`Mycobacterium S.`, `Mycobacterium T.`}. Le commentaire de l'expert était de dire que la distinction entre `Neutral Gram` et `Positive Gram` n'avait pas de sens et qu'il fallait regrouper ces deux attributs en un seul. La figure 3.3 donne une vue globale du treillis et des concepts possédant ces deux propriétés.

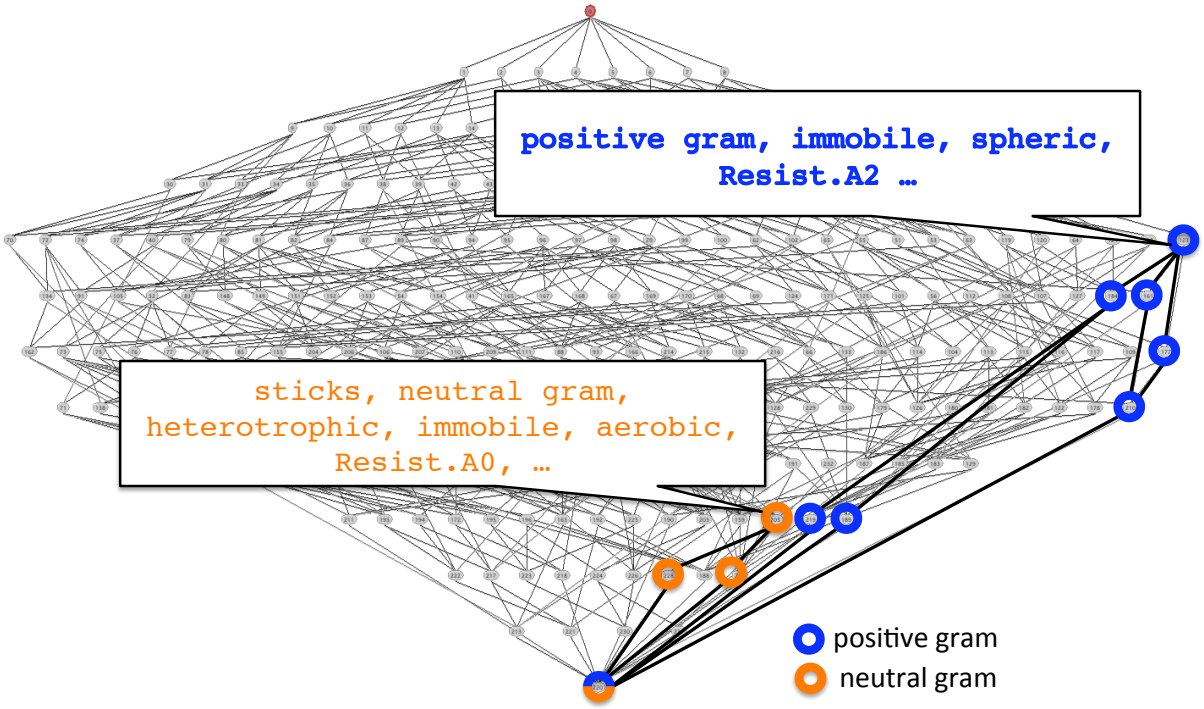


FIGURE 3.3 – Concepts du treillis impliquant les propriétés GramPositif ou GramNeutre

La figure 3.4 est le treillis après regroupement de la propriété GramNeutre avec l'attribut GramPositif. Sans entrer dans le détail de la différence entre les concepts des deux figures, on observe que la fusion des deux propriétés peut engendrer de nouveaux concepts dans le treillis (ou en supprimer) et modifie ainsi de façon assez profonde la conceptualisation proposée à l'expert.

Cela nous a conduit à définir un ensemble d'opérations pouvant être proposées à l'expert pour interagir sur les données

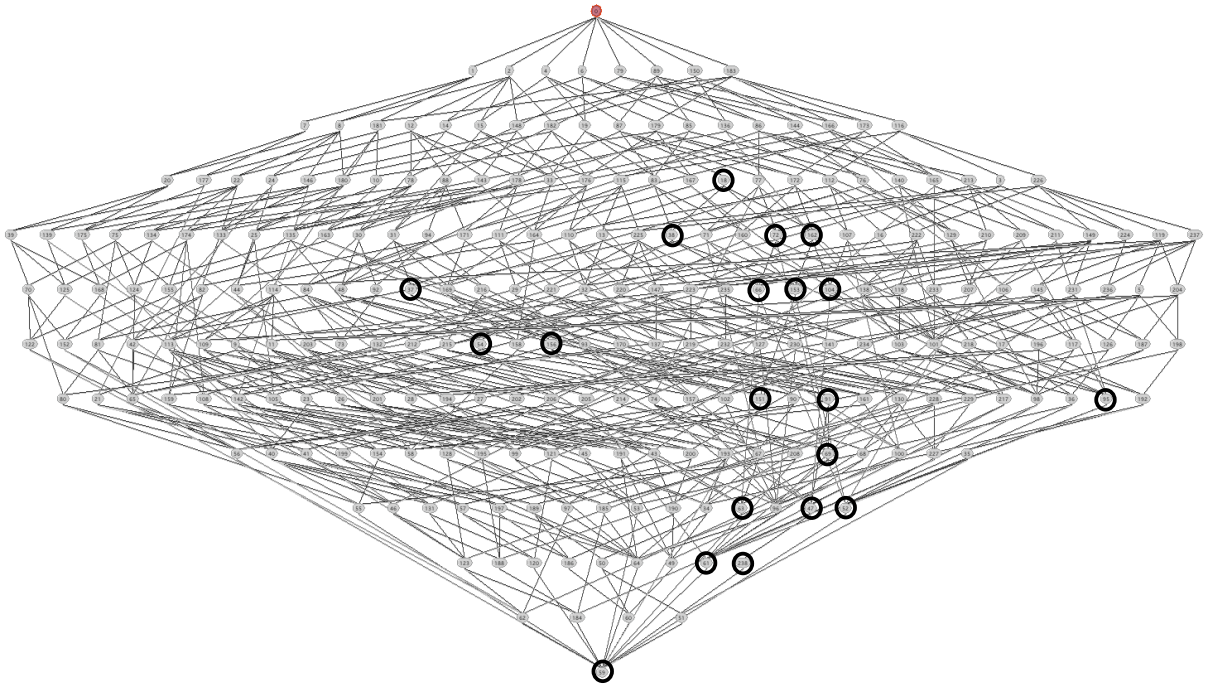


FIGURE 3.4 – Concepts du treillis impliquant les propriétés Gram Positif ou Gram Neutre

3.5 Conclusion sur la construction d'ontologie à partir de textes

LA FCA apporte une réponse formelle à la construction de concepts et la hiérarchisation de ces concepts lors de la construction d'ontologie. Sans cette aide, c'est une étape complexe à mener à la main par un expert, même s'il se place dans un environnement comme PROTÉGÉ [NFM00] couplé à un raisonneur comme PELLET [PS04].

L'ontologie est construite par une démarche "bottom-up" à partir de la description d'individus et de leurs attributs. L'analyse relationnelle des concepts, en tant qu'extension à la FCA, permet de plus de prendre en compte plusieurs contextes formels et les relations binaires intra ou extra contextes entre les individus.

Les propriétés mathématiques des treillis confèrent aux concepts des propriétés définitoires qui sont directement exploitées pour coder l'ontologie en logique de descriptions. De plus, les opérations proposées par la FCA, comme par exemple l'apposition, permettent également d'intégrer des connaissances éventuellement extérieures aux textes telle que des classes d'objets prédéfinies issues de thésaurus.

Enfin, un des points forts de cette approche réside dans sa capacité à proposer des définitions aux classes (ou concepts). C'est une tâche souvent négligée dans une ontologie dans laquelle une attention plus particulière est généralement portée à la construction de la hiérarchie de concepts. L'expérience en astronomie a montré qu'il était possible de partir d'un thésaurus et de produire des définitions pour une certain nombre des catégories du thésaurus, parfois par des conditions nécessaires, parfois par des conditions nécessaires et suffisantes. Le même processus propose également à l'expert de nouvelles classes qui

raffinent les classes du thésaurus. Nous avons par exemple exploité ces propriétés dans le cadre de la réorganisation ou restructuration (re-engineering) d'un wiki sémantique [STNB11].

Intégrer la vision de l'expert reste un point difficile que nous aborderons dans la section 5.4.1. L'expert ne se place généralement pas au niveau de l'objet mais adopte plutôt une démarche top-down. Le treillis doit être alors vu comme l'espace des concepts possibles, laissant à l'expert la tâche de déterminer ceux qui sont les plus pertinents pour son domaine.

Il reste de nombreux travaux à faire sur la construction de l'ontologie. Si la FCA s'avère être un formalisme très puissant, les expérimentations actuelles restent très localisées. Seule une toute petite partie de l'information des textes est extraite et exploitée pour construire les contextes formels, que ce soit des informations de type attribut ou relationnel. C'est donc une question sur le passage à l'échelle d'une telle approche et, nous le verrons en section 4 des défis pour la préparation des textes.

Chapitre 4

Synthèse sur la préparation des documents

Comme le montre le schéma sur l'extraction de connaissances 1.1, la fouille de données intervient après une étape essentielle de préparation des données. [BCM05] est une collection d'articles éclairant les différentes étapes du processus. Nous abordons donc dans ce chapitre les questions liées à la préparation des textes. Ces questions s'inscrivent dans la problématique du traitement automatique de la langue (TAL), mais avec une visée partielle et robuste. Compte tenu de la diversité des approches, du coût humain et en temps très important pour la mise en adéquation de ces méthodes à un domaine particulier, il faut voir les travaux et les questions que je soulève dans ce chapitre comme autant de collaborations possibles avec des équipes dont c'est le coeur d'activité.

Exploité par les travaux en psycholinguistique [Bib92], les premières connaissances accessibles à partir des textes sont les connaissances linguistiques. Hearst [Hea04] souligne cependant que le résultat d'un processus de fouille de textes ne peut se limiter à l'observation de cooccurrences et que les observations de nature linguistique ne suffisent pas aux experts en attente de connaissances sur leur domaine :

« If we extrapolate from data mining (as practiced) on numerical data to data mining from text collections, we discover that there already exists a field engaged in text data mining : corpus-based computational linguistics ! Empirical computational linguistics computes statistics over large text collections in order to discover useful patterns. These patterns are used to inform algorithms for various subproblems within natural language processing, such as part-of-speech tagging, word sense disambiguation, and bilingual dictionary creation ([Arm94]).

It is certainly of interest to a computational linguist that the words "prices, prescription, and patent" are highly likely to co-occur with the medical sense of "drug" while "abuse, paraphernalia, and illicit" are likely to co-occur with the illegal drug sense of this word ([CL91]). This kind of information can also be used to improve information retrieval algorithms. However, the kinds of patterns found and used in computational linguistics are not likely to be what the general business community hopes for when they use the term text data mining. » [Hea04]

4.1 Constitution de corpus

4.1.1 Collecter les textes

La constitution du corpus pour la fouille de textes est une étape essentielle dans laquelle les pratiques et les méthodes sont aussi importantes que les outils. Par outils, nous entendons bien sûr, les travaux classiques issus de la recherche d'information pour l'interrogation de bases documentaires et la sélection de documents mais aussi les différentes méthodes de catégorisation, de classification et de cartographie qui sont toujours très utilisées pour découvrir des tendances dans certains domaines. L'approche statistique dite « des mots associés » a permis la réalisation d'outils comme LEXIMAPPE ou SDOC alors que d'autres approches exploitent des techniques neuronales [Koh84, LTS03].

Une caractéristique commune à la plupart de ces travaux est la robustesse des techniques mises en œuvre et leur capacité à traiter de très gros volumes de données ou de textes. L'analyse se fait à un niveau très superficiel. L'unité couramment exploitée est le mot ou le mot-clé et la notion de sens n'est pas vraiment pertinente au profit de la notion de pouvoir discriminant.

La classification peut être vue comme un moyen d'augmenter la cohésion d'un corpus. Ainsi, d'un corpus global sur l'agriculture, on peut distinguer des sous-ensembles de textes traitant plus particulièrement des questions liées à la croissance du maïs ou liées à la conservation et au transport du maïs. De même, en bibliométrie, la classification construit des réseaux d'auteurs, d'institutions ou de manifestations scientifiques qui peuvent être exploités pour contraindre ou augmenter un ensemble de textes initialement constitué.

Ce niveau d'analyse, bien qu'assez pauvre du point de vue sémantique permet néanmoins de structurer un domaine. Ces techniques sont utilisées en veille technologique ou en intelligence économique. Elles sont peu coûteuses à mettre en œuvre et polyvalentes. Certaines méthodes, comme les cartes de Kohonen multi points de vue [LTS03] sont appréciées pour leur lisibilité. L'expert en charge de l'analyse se sert alors de classes comme d'un support pour formuler des phrases explicatives, voire des hypothèses qu'il pourra vérifier en accédant directement aux textes. L'échelle – le nombre de documents pris en compte et le nombre de classes générées – conditionne bien évidemment la qualité de cette analyse.

L'homogénéïté d'un corpus est un atout mais également un écueil. Selon les types de textes, les informations, voire même les connaissances qui en sont extraites, sont de natures différentes. Les textes didactiques, comme par exemple le précis de Géomorphologie déjà cité précédemment, contiennent les définitions des notions importantes et bien établies dans un domaine mais, en contrepartie, ces informations sont mises à jour avec une périodicité longue, lors d'une nouvelle édition ou d'une nouvelle parution. Les articles scientifiques font, quant à eux, référence à de nombreuses connaissances implicites mais s'inscrivent dans un monde en évolution permanente : les objets manipulés dans ces textes sont rarement associés à leurs hyperonymes ou à leurs catégories, cela fait partie de la connaissance commune. La terminologie peut être fluctuante, les conditions d'expérimentation très diverses et les résultats rapportés peuvent également être contradictoires. Les thésaurus, bien que construits de façon informelle, structurent un domaine par exemple, en introduisant des catégories et une hiérarchie entre ces catégories. Ces différentes res-

sources sont donc complémentaires les unes des autres et posent des défis spécifiques en matière de prétraitement. Ainsi, extraire les catégories d'un thésaurus peut être peaufiné et "nettoyé" à la main, même s'il s'agit de 5 000 entrées, pour peu qu'on ne le fasse qu'une fois. En revanche, traiter 80 000 textes ou un flux variable de textes nécessite une procédure entièrement automatisée et, bien souvent, capable d'évoluer en fonction des retours d'évaluation.

Il faut ensuite délimiter les portions de textes pertinents pour le problème traité et les approches statistiques donnent de bons résultats : les SVM, la classification naïve bayésienne (CNB), les réseaux de neurones ou les arbres de décision. Selon [NVB01], la CNB donne les meilleurs résultats avec 90% de précision et de rappel et l'apport d'un prétraitement linguistique (lemmatisation, terminologie) ne semble pas beaucoup influencer ces résultats.

4.1.2 Quelle information extraire des textes ?

Les types de textes que nous venons d'introduire ne suffisent pas encore à déterminer les traitements qui doivent leur être appliqués. Si l'on fait abstraction des problèmes liés au choix de la langue ou au multilinguisme que je n'aborderai pas dans ce mémoire, il n'en demeure pas moins que les textes d'un même domaine contiennent des informations de natures très différentes. Les figures 4.1, 4.2 et 4.3 donnent l'exemple de trois résumés d'articles scientifiques en pharmacovigilance issus de PubMed.

Le texte en figure 4.1 décrit ce qu'on appelle un cas (en particulier) en pharmacovigilance, c'est-à-dire la situation d'un patient à qui de l'héparine a été injectée et qui a fait une thrombocytopénie. Le résumé décrit les caractéristiques du patient (femme de 29 ans, en surpoids, ayant déjà suivi un traitement à l'enoxaparine) ainsi que les différents tests qui ont été appliqués.

Le texte en figure 4.2, en revanche, dresse un bilan des observations réalisées sur 10 ans sur la thrombocytopénie immunologique induite à l'héparine. C'est un des rares effets indésirables médicamenteux pour lequel il existe une physiopathologie.

Enfin, les essais cliniques ou l'analyse comparée de plusieurs cas (figure 4.3) rapportent différentes expérimentations d'un même produit, parfois administré sous des formes différentes et des populations différentes dans le cas d'essais cliniques. Le bilan est alors exposé par rapport aux paramètres produit, démographique... Ce dernier exemple montre bien que la présence de trois termes – médicament, personne et effet indésirable – ne suffit pas à conclure qu'il s'agit d'un cas en pharmacovigilance. On peut d'ailleurs observer combien il est difficile de synthétiser, même à la main, la situation de chacun des patients.

4.2 Les bases de connaissances terminologiques

Au début des années 1990, l'article de [SM91] a été repris en France par le groupe Terminologie et intelligence artificielle. Il s'est alors créé un véritable rapprochement entre la terminologie et l'intelligence artificielle autour de la notion de base de connaissances

Difficulties in the detection of heparin-induced thrombocytopenia type II
[Schweiz Rundsch Med Prax, 1999]
We report about a 29 year old female who developed right-sided leg vein thrombosis over three levels. Thrombectomy was attempted followed by intravenous anticoagulation with heparin. The platelet count dropped acutely from 176,000/microliter to 11,000/microliter after the sixth day. A lung perfusion-ventilation-scintigraphy suggested recent pulmonary embolism by lateral, predominantly right-sided perfusion deficits. ACT scan of the pelvic region showed rethrombosis of the right common iliac vein. The clinical suspicion of heparin-induced thrombocytopenia (HIT) type II was confirmed by a positive heparin-induced platelet aggregation test and the detection of antibodies by heparin-platelet factor 4-ELISA. The patient was treated with lepirudin at body-weight-adapted dose. After recovery of the platelet count to 102,000/microliter within seven days the treatment was changed to Orgaran after exclusion of immunologic cross reactivity. An overlapping oral anticoagulation with Marcoumar was initiated. Although HIT type II usually develops over a few days, acute thrombopenia can also occur. There is therefore no safe diagnostic interval permitting a timely detection.

FIGURE 4.1 – Exemples d'un cas en pharmacovigilance (extrait de PubMed)

Heparin-induced thrombocytopenia : a ten-year retrospective
[Annu Rev Med 1999]
The past decade has seen many important advances in the pathogenesis, clinical and laboratory diagnosis, and management of heparin-induced thrombocytopenia (HIT), one of the most common immune-mediated adverse drug reactions. HIT is caused by IgG antibodies that recognize complexes of heparin and platelet factor 4, leading to platelet activation via platelet Fc gamma I Ia receptors. Formation of procoagulant, platelet-derived microparticles, and, possibly, activation of endothelium generate thrombin in vivo. Thrombin generation helps to explain the strong association between HIT and thrombosis, including the newly recognized syndrome of warfarin-induced venous limb gangrene. This syndrome occurs when acquired protein C deficiency during warfarin treatment of HIT and deep venous thrombosis leads to the inability to regulate thrombin generation in the microvasculature. The central role of HIT antibodies in causing HIT, as well as refinements in laboratory assays to detect these antibodies, means that HIT should be considered a clinicopathologic syndrome. The diagnosis can be made confidently when one or more typical clinical events (most frequently, thrombocytopenia with or without thrombosis) occur in a patient with detectable HIT antibodies. The central role of thrombin generation in this syndrome provides a rationale for the use of anticoagulants that reduce thrombin generation (danaparoid) or inhibit thrombin (lepirudin).

FIGURE 4.2 – Textes de synthèse en pharmacovigilance (extrait de PubMed)

Heparin-induced vascular occlusion in vasculosurgical patients
[J. Cardiovasc Surg (Torino), 1999]
An evaluation of the disease in 13 cases. OBJECTIVE : To describe the diagnosis and treatment of adverse reaction to heparin (heparin-induced thrombocytopenia [HIT]) administered prophylactically for thrombosis and embolism. Experimental design : case series. Setting : vascular surgical division in a University hospital. Patients : thirteen patients treated for HIT type II between October 1994 and June 1997. Measures/Interventions : diagnosis of heparin-induced complications is based on exact medical history and regular measurement of platelet counts. Confirmation can be obtained with the aggregation test, serotonin-release test, heparininduced platelet release (HIPA) test, and platelet factor 4/heparin ELISA. Vasculosurgical reconstruction is usually required to eliminate vessel occlusion. RESULTS : In our series, HIT was confirmed by HIPA test (11 patients) and aggregation test (2 patients). All patients had positive cross reaction with low-molecular-weight heparin, and six had cross reaction with heparinoid danaproid sodium (Orgaran). Occlusions occurred between day 2 and 22 after the start or resumption of heparin administration (mean, 11 days). Anticoagulation treatment with hirudin or danaproid sodium was given to 5 patients, in conjunction with vasculosurgical reconstruction. Three of those patients died and the other two required amputation. CONCLUSIONS : Heparin-induced vascular occlusion is a rare but severe adverse effect of heparin treatment. When HIT is suspected, heparin administration must be stopped, with substitution of dextran and acetylsalicylic acid. Laboratory tests must be used for confirmation or exclusion. However, the diagnosis can be obscured by a normal platelet count due to pre-existing polycythemia and by false-negative test results. Surgery is usually warranted, depending on the degree and localization of ischemia.

FIGURE 4.3 – Analyse comparée de plusieurs cas en pharmacovigilance (extrait de PubMed)

terminologiques [4]. Du côté de l'intelligence artificielle, deux communautés se sont plus particulièrement impliquées : d'une part l'ingénierie des connaissances qui s'intéresse tout particulièrement à la construction d'ontologies et, d'autre part, le traitement automatique de la langue (TAL) qui recherche à travers la terminologie le moyen de relier les mots à des concepts et qui apporte également des outils de traitement partiels et robustes. Du point de vue de la terminologie, ce rapprochement a entraîné une réflexion sur la construction du sens d'un texte dans laquelle la vision purement référentielle était inappropriée pour lui préférer une sémantique textuelle [Con03].

Cette réflexion autour de la terminologie a coïncidé avec un regain d'intérêt pour la linguistique de corpus et l'arrivée d'outils de traitement de corpus performants, notamment, les étiqueteurs.

4.2.1 Etiquetage morpho-syntaxique et lemmatisation

L'étiquetage des textes L'étiquetage morpho-syntaxique consiste à assigner à chaque mot d'un texte sa catégorie morpho-syntaxique (nom, verbe, adjectif...). L'intérêt de l'étiquetage est qu'il rend possible l'analyse d'un corpus du point de vue linguistique et plus statistique (basé sur les fréquences d'association). A partir d'un texte étiqueté, il est en effet possible de caractériser ou de rechercher certaines structures syntaxiques, c'est-à-dire, de s'abstraire du mot pour passer à sa catégorie linguistique. Par exemple, la phrase (1) est étiquetée par (2) :

1. Two resistant strains were isolated after four rounds of selection.
2. Two/CD resistant/JJ strains/NNS :pl were/VBD isolated/VBN after/IN four/CD rounds/NNS :pl of/IN selection/NN ./.

Il existe plusieurs étiqueteurs morpho-syntaxiques pour le français, l'anglais et également pour d'autres langues européennes. D'autres outils, comme les analyseurs syntaxiques partiels (*chunker*) intègrent parfois leur propre étiqueteur.

L'étiqueteur de Brill [Bri92, Bri93] ou le TreeTagger [Sch94] sont des outils très couramment utilisés dans la communauté francophone. Suivant la nature du corpus et le domaine, les performances de ces outils vont de 95 % à 99 %. Ces outils exploitent des techniques d'apprentissage et leur performance dépend du jeu d'étiquettes, de la qualité et de la quantité du corpus d'entraînement. L'étiqueteur de Brill peut être complètement réentraîné (l'INaLF [5] l'a ainsi adapté au français) alors qu'il n'est possible que de compléter l'entraînement du TreeTagger pour l'adapter, par exemple à un nouveau domaine. Il faut souligner que ces outils sont robustes et qu'ils fournissent donc une réponse, même pour les mots qu'ils n'ont jamais rencontrés. Brill, par exemple, opère en utilisant des règles lexicales qui exploitent la morphologie du mot pour en prédire la catégorie.

Les étiqueteurs ne sont pas tant sensibles aux mots inconnus liés à un domaine particulier qu'aux structures syntaxiques spécifiques à un domaine. Les règles de Brill permettent de prédire la catégorie d'un mot inconnu. En revanche, appliqué à un corpus de recettes de

4. Ces travaux ont été menés par le groupe terminologie et intelligence artificielle (TIA) du GDR-PRC I3.

5. Institut national de la langue française qui est devenu maintenant sur Nancy l'ATILF (Analyse et traitement informatique de la langue française).

cuisine en anglais (projet Taaable que nous décrivons en section 5.4) composé quasi exclusivement de phrases impératives et contenant beaucoup de mots ambigus (mots pouvant être soit un verbe soit un nom), le taux de mots correctement étiquetés chute en dessous de 95%.

La lemmatisation L'étiquetage est souvent à la base de tous les travaux en linguistique de corpus mais il n'apporte pas vraiment de transformation à l'énoncé initial. Un certain nombre d'autres travaux permettent non seulement de repérer certains phénomènes mais aussi d'en effectuer des regroupements motivés linguistiquement. Par rapport à la fouille de textes, l'idée est de pouvoir opérer des regroupements sémantiquement interprétables. La lemmatisation permet de ramener les mots fléchis à une forme canonique [Nam00]. La morphologie dérivationnelle qui s'introduit progressivement dans le traitement de la terminologie puis de la fouille de textes vise à identifier un sens construit à partir de la base d'un mot et d'un ensemble de suffixes et à proposer une glose qui, par la suite pourrait être formalisée. Par exemple, le système DeriF [Nam04] produit : benladenisation, NOM → [[[Benladen NPR] is(er) VERBE] tion NOM] (benladenisation/NOM, benladeniser/VERBE, Benladen/NPR) : : « action ou résultat de benladeniser ».

4.2.2 L'extraction terminologique

Un terme est une unité syntaxique composée d'un mot ou d'un groupe de mots considérée comme une notion importante dans le domaine de spécialité étudié et qui peut être associée à un concept dans une base de connaissances.

Il existe plusieurs types d'outils pour extraire une terminologie d'un corpus et le choix d'un outil en particulier doit s'effectuer en lien avec la finalité du processus de fouille. Au même titre que l'on distingue l'indexation libre de l'indexation contrôlée, nous distinguons l'extraction terminologique libre de l'extraction contrôlée. Il faut souligner qu'aucune de ces méthodes ne permet de construire une base de connaissances ou une ontologie. Ce n'est là qu'une des étapes du processus global.

L'extraction terminologique libre. L'extraction terminologique libre peut se faire suivant une stratégie ascendante ou descendante. La stratégie ascendante consiste à rechercher dans les textes des collocations entre mots, à les organiser suivant des indices statistiques pour identifier ceux qui seraient potentiellement des termes. C'est une approche robuste dans la mesure où il est possible de rechercher les collocations de mots, de mots étiquetés ou de mots étiquetés et lemmatisés suivant la nature des prétraitements effectués. Elle peut, de ce fait, être peu dépendante de la langue. ANA [EP95] a été développé suivant ce principe. L'approche descendante se base sur des patrons linguistiques, notamment la synapsie [Ben66] qui peuvent potentiellement délimiter des termes. Ainsi Nomino [DP90] travaille-t-il à partir de patrons du type *Nom Prep Nom*, *Nom* de *Nom* ou encore *Nom Adjectif*. Enfin, plusieurs travaux combinent une approche descendante avec une approche ascendante, ou réciproquement. Il s'agit par exemple des outils ACABIT [Dai94], Lexter [Bou94] ou Xtract [Sma93]. Dans tous les cas, l'extraction libre ne donne

pas des termes mais des candidats-termes : ces outils privilégient, en effet, un silence faible (peu de termes qui auraient dû être proposés sont oubliés) au détriment du bruit qui est assez élevé (beaucoup de candidats termes ne sont pas des termes).

L'extraction terminologique contrôlée. Le principe de l'extraction terminologique contrôlée est de chercher à reconnaître les termes appartenant à un référentiel connu : thésaurus ou simple liste de termes. Contrairement à l'extraction libre, cette approche génère assez peu de bruit mais plus de silence. Pour réduire ce silence et rechercher des termes qui n'apparaissent pas toujours sous une forme strictement identique à la forme initiale, les travaux de [JR94] exploitent la notion de variante linguistique des termes. Cette notion de variante a d'ailleurs également été intégrée dans l'extraction libre. L'idée est que les termes apparaissant sous des formes variantes désignent le même concept. On cherche donc des variantes qui préservent le sens du terme initial. L'expression comme « acte de terrorisme » est ainsi considérée comme une variante du terme « acte terroriste ». Pour préserver le sens, FASTR (outil de reconnaissance des termes et de leurs variantes) fait appel à des métarègles de transformations linguistiques qui supposent que le corpus initial soit étiqueté et lemmatisé. Certaines dérivations morphologiques peuvent aussi être prises en compte ainsi qu'un certain typage sémantique, mais cela dépend de la qualité de la phase de préparation des textes. On se réfèrera à [Jac01] ou à [Dai02] pour une analyse plus détaillée de la notion de variation et des différents types de variations. La variation, suivant le domaine et le type de corpus représente entre 15 et 30% des occurrences des termes.

Le terme est une unité peu exploitée en recherche d'information. D'une part, le coût de préparation des textes est plus élevé que pour le mot. D'autre part, la prise en compte des termes au lieu des mots engendre une baisse des fréquences et suppose de travailler dans un espace de plus grande dimension. Le gain est faible, voire négatif pour la recherche d'information. En fouille de textes, le travail au niveau du terme peut être très positif. Il est bien sûr nécessaire d'identifier les termes parmi l'ensemble des candidats-termes, ce qui est réalisé généralement, à la main, par un expert. L'occurrence des termes dans des classes est beaucoup plus significative et précise que l'occurrence des mots. Dans le cas de la veille technologique, par exemple, l'expert est mieux à même de verbaliser le contenu des classes. De plus, la prise en compte des formes variantes des termes et le regroupement des variantes autour de la forme initiale réduit la dispersion, et réduit donc la complexité des processus de fouille, que ce soit en classification ou en extraction de règles d'association tant sur le plan du calcul que sur le plan de l'interprétation par un expert. L'extraction des termes et de leurs variantes étaient à la base des travaux que j'ai menés en lien avec l'INIST [6] dans le cadre de la plateforme de veille technologique ILC (Informatique, Langage & Connaisssances).

Il est intéressant de rappeler des résultats assez anciens d'une expérimentation [DRP00] sur un corpus de 2 702 notices bibliographiques dans le domaine de l'agro-alimentaire visant à évaluer 3 volets : l'indexation contrôlée, l'indexation libre et la pertinence des variantes identifiées lors de l'indexation contrôlée. La table 4.1 souligne tout particulièrement l'importance du phénomène de variation en corpus :

6. Institut de l'information scientifique et technique, UPS 0076 du CNRS.

TABLE 4.1 – Proportion des phénomènes de variation sur le corpus AGRO-ALIM identifiés par FASTER

Termes	**Variantes syntaxiques**			**Variantes morpho-syntaxiques**					
	Coord	Modif	Synap	NàV	NàN	NàA	AàN	AàV	AàAv
55812	2165	9983	4216	7259	9213	2663	1414	90	2
55812 (60,1%)	16364 (17,6%)			20641 (22,3%)					

L'extraction de relations entre termes Les relations entre les termes sont très importantes pour la construction d'une base de connaissances. Les systèmes SEEK [Jou93] ou COATIS [Gar98] travaillent à partir de marqueurs de relations définis *a priori*. Les travaux comme Promethee [Mor99], se rapprochent davantage de techniques d'apprentissage : si une relation ρ a été identifiée entre deux termes $\mathtt{t}_1$ et $\mathtt{t}_2$, la recherche de toutes les phrases où ces deux termes apparaissent permet alors d'identifier les différents marqueurs de cette relation. Dans un contexte un peu différent, [Seb02] utilise la programmation logique inductive pour l'apprentissage de relations lexicales et leur caractérisation.

L'extraction de relations entre termes peut être rapprochée des phases d'apprentissage en extraction d'information (voir section 4.3). En effet, tous ces travaux identifient des notions puis cherchent des marqueurs permettant de les retrouver en corpus. Nous abordons cette question dans la section 4.3.

Les limites de la représentation par les termes L'extraction terminologique peut être considérée actuellement comme un processus robuste ayant acquis une bonne maturité. Cependant, du point de vue de la fouille de textes, un certain nombre de difficultés subsistent. La sélection des termes parmi les termes-candidats nécessite un travail important. L'extraction de relations est également un travail long qui reste *ad hoc* et dépendant du domaine.

Par ailleurs, la représentation d'un texte par un ensemble de termes, tout comme sa représentation par un sac de mots, est trop pauvre. Ainsi, dans une phrase comme :
« When maintained under nonselective conditions, neither the aadA mRNA nor the AadA protein were detected in these subclones »
le terme mRNA sera identifié au même titre que dans la phrase
« Identification of an ABC transporter gene that exhibits mRNA level overexpression in fluoroquinolone-resistant Mycobacterium smegmatis ».
Pourtant, dans la première phrase, la négation montre bien que mRNA n'est pas présent dans cette expérience alors qu'il l'est dans la seconde. La représentation par un ensemble de termes ne peut traiter la négation ce qui peut, par la suite, engendrer des erreurs de classification et d'interprétation.

Enfin, l'indexation contrôlée ne permet pas de repérer de nouveaux termes et limite donc la portée de la veille à des notions déjà connues.

4.3 L'extraction d'information

Le traitement automatique de la langue, comme je l'ai souligné en section 1.3, cherche en partie à modéliser le fonctionnement de la langue et il est donc naturellement inspiré des travaux en linguistique : construction de grammaires formelles, de lexiques, construction du sens (en logique) d'une phrase ou d'un discours... L'analyse d'un énoncé peut se faire en profondeur créant un arbre syntaxique complet où chaque mot est analysé. C'est sur cette structure syntaxique qu'est construite une représentation sémantique. L'analyse peut également être partielle, analysant localement des parties de la phrase. Dans ce cas, les outils peuvent exploiter des principes linguistiques ou statistiques, voire les deux, tel que cela est couramment utilisé pour la segmentation de textes (en thème ou en phrase), pour l'étiquetage morpho-syntaxique, pour la désambiguïsation des mots, le rattachement prépositionnel ou la résolution d'anaphore... Ces travaux sont souvent orientés vers une "compréhension fine" de l'énoncé, et les processus sont trop complexes pour traiter de grandes quantités de textes réels avec toute la diversité dans les phénomènes linguistiques que cela implique.

La représentation logique, dans la lignée des travaux de la sémantique de Montague [Mon73, Gam91] peut être exploitée par des démonstrateurs logiques dans lesquels il est possible d'introduire des connaissances du domaine. Il est cependant difficile de s'abstraire du niveau linguistique pour passer à un niveau conceptuel. Nous citons en section 1.1.3 l'exemple des deux phrases *« Heat ghee in a large soup pot»* et *« Melt butter in a large stock pot over a medium heat»* extraites de recettes de cuisine. Les deux phrases font référence à la chaleur ("heat") mais dans la première phrase, ce mot est un verbe que l'on peut donc représenter sémantiquement comme une structure prédicative alors que dans la seconde phrase, c'est un nom utilisé pour désigner la nature du feu. La première phrase fait chauffer de la graisse tandis que la seconde fait fondre du beurre. Pourtant, le résultat de ces deux actions est, dans les deux cas, de la matière grasse chaude et liquide dans une grande casserole.

Au lieu de partir de la forme linguistique pour aller vers le conceptuel, l'Extraction d'Information (EI), en quelque sorte, fonctionne dans l'autre sens. Le point de départ est un ensemble de structures conceptuelles ou de frames pour lesquels on recherche les marqueurs linguistiques qui vont permettre de les identifier dans les textes. De ce fait, elle exploite des techniques issues de la recherche d'information et de l'apprentissage automatique. Aujourd'hui, si les travaux en extraction d'information figurent dans les grandes conférences de traitement automatique de la langue aux côtés des travaux d'inspiration plus logique et linguistique, l'intégration n'est pas encore parfaite. Mais, de ce point de vue, les perspectives semblent variées et très riches.

Le principe sous-jacent à l'extraction d'information (EI) est de décomposer l'identification d'informations, parfois complexes, en des sous-problèmes simples. Ainsi, les grandes tâches classiquement identifiées en EI sont [HDG00] :

- reconnaissances d'entités nommées : noms de personnes, d'organisations, expressions temporelles comme les dates ou les heures, unités de mesures...
- résolution des coréférences : pronoms, la désignation par la fonction (Premier Ministre), abréviations et variantes orthographiques (béta-lactamase, $\beta-$lactamase)...
- extraction de propriétés : localisation géographique, association entre une personne

et ses fonctions...
- identification de relations : interaction entre des protéines, personne employée par une société...
- identification des événements : les événements sont des relations complexes telle que la description d'une attaque terroriste dans laquelle doivent être identifiés les terroristes, les victimes, les lieux, la date...

Les objectifs de l'EI consistent à remplir automatiquement un formulaire à partir d'un énoncé en langue naturelle. Il ne s'agit pas uniquement de filtrer des parties d'un énoncé car comme le montre [Ned04] dans l'exemple repris en Figure 4.4, remplir des champs nécessite souvent la prise en compte du contexte, mais aussi des types d'objets manipulés ou du type de textes...

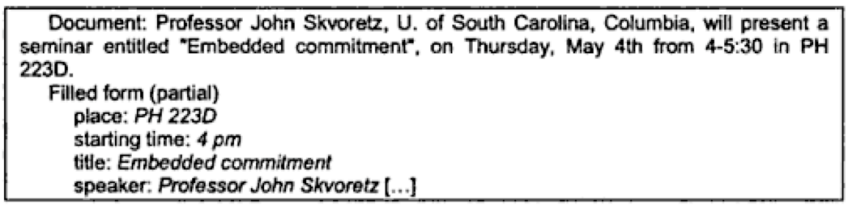

Document: Professor John Skvoretz, U. of South Carolina, Columbia, will present a seminar entitled "Embedded commitment", on Thursday, May 4th from 4-5:30 in PH 223D.
Filled form (partial)
place: *PH 223D*
starting time: *4 pm*
title: *Embedded commitment*
speaker: *Professor John Skvoretz* [...]

FIGURE 4.4 – Exemple d'extraction d'information sur l'annonce d'un séminaire

Les sous-sections suivantes se focalisent sur quelques tâches réalisées en extraction d'information et les méthodes utilisées. Un certain nombre d'exemples, de résultats d'évaluation et de références sont empruntés à [MB05, KSS08] ou à [Ned04] qui propose un état de l'art approfondi dans le cadre de la génomique. Les perspectives qui m'intéressent, dans mon projet de recherche, sont bien identifiées dans [Ned04] et reprises et développées dans [NN05] dans lequel les liens avec les ontologies sont mis en avant.

4.3.1 Quelques éléments en extraction d'information

Reconnaissance d'entités nommées

Il existe de très nombreux travaux sur l'apprentissage statistique pour l'extraction d'information à des domaines également très divers.

La reconnaissance d'entités nommées a pour but d'identifier dans les textes les objets spécifiques comme des personnes, des noms de sociétés ou des noms de lieux mais également, des noms de protéines, de gènes... Des ressources propres au domaine peuvent être exploitées : des thésaurus, des dictionnaires ou encore de patrons comme cela peut-être le cas pour repérer des galaxies en astronomie (*NGC xxxx*, où *xxxx* sont des chiffres).

[Ned04] montre que les approches les plus répandues pour l'identification d'entités nommées reposent sur des méthodes d'apprentissage numériques. La modélisation statistique de séquences par les Modèles de Markov Cachés (HMM) [Rab02] ou les CRF (Conditional Random Fields) qui semblent un peu plus performants [LMP01, PM04, SC05], font partie des méthodes très utilisées pour le repérage des entités nommées, notamment dans le system Nymble [BMSW97]. L'apprentissage est effectué sur un corpus annoté et le système annote un nouveau document en assignant l'étiquette la plus probable. D'autres

Source :
SOCCER - BLINKER BAN LIFTED.
LONDON 1996-12-06 Dutch forward Reggie Blinker had his indefinite suspension lifted by FIFA on Friday and was set to make his Sheffield Wednesday comeback against Liverpool on Saturday. Blinker missed his club's last two games after FIFA slapped a worldwide ban on him for appearing to sign contracts for both Wednesday and Udinese while he was playing for Feyenoord.

Result :
SOCCER - [PER BLINKER] BAN LIFTED . [LOC LONDON] 1996-12-06 [MISC Dutch] forward [PER Reggie Blinker] had his indefinite suspension lifted by [ORG FIFA] on Friday and was set to make his [ORG Sheffield Wednesday] comeback against [ORG Liverpool] on Saturday . [PER Blinker] missed his club 's last two games after [ORG FIFA] slapped a worldwide ban on him for appearing to sign contracts for both Wednesday and [ORG Udinese] while he was playing for [ORG Feyenoord] .

FIGURE 4.5 – exemple d'extraction d'entités nommées dans un texte (Illinois NER).

approches encore reposent sur le principe du maximum d'entropie [Mik98] ou sur les SVM [TC02].

Certains de ces outils sont accessibles, comme par exemple, le Stanford NER ou l'Illinois NER qui est illustré par une démonstration en ligne (voir figure 4.5).

L'utilisation des règles d'association séquentielles pour l'extraction d'entités nommées semble une alternative intéressante [PCK+09]. Nous reviendrons sur l'intérêt de ces travaux dans le cadre de mon projet de recherche (section 5) dans la mesure où ils reposent sur les mêmes fondements scientifiques que ceux utilisés pour la construction de l'ontologie.

Extraction de règles

Le remplissage des champs d'un formulaire nécessite d'aller au delà de l'identification des termes ou des entités nommées puisqu'il s'agit de relier un champ à un contenu, ou plus globalement d'identifier des structures de type prédicat-argument, ce qui revient à extraire des relations entre des objets du domaine.

L'extraction de relations peut se faire par la définition, puis la recherche dans les textes, de patrons, impliquant des objets (termes), des entités nommées, et des verbes (ou d'autres marqueurs linguistiques) qui reflètent une relation. Les règles peuvent aussi prendre en compte des contraintes de nature syntaxique. Dans le cadre de la génomique, [OHTT01] repose ainsi sur l'utilisation d'un dictionnaire pour les noms de protéines, de patrons de mots et d'un étiquetage morpho-syntaxique. La figure 4.6 donne un exemple de patrons définis pour la reconnaissance d'interaction entre protéines.

Dans ces travaux, deux types de règles sont définies : les règles positives qui correspondent à des situations d'interaction à observer et des règles négatives qui permettent de filtrer les situations que l'on sait ne pas être une interaction. L'idée derrière ces règles négatives est de réduire le nombre de faux positifs dans l'extraction. Un exemple de règle est *PROTEIN1.*PATTERN.* but NOT PROTEIN2* dans laquelle *PATTERN* est une forme

Keyword	Pattern	Example of sentence
Interact	*A* interact with *B*	*Spc97p interacts with spc98* and *Tub4* in the two-hybrid system.
	interaction of *A* (with|and) *B*	The *interaction of Cet1 with Ceg1* elicits...
	interaction (between|among) *A* and *B*	Functional and physical *interaction* between *Rad24* and *Rfc5*...
	A–*B* interaction	These data suggest that the *Cert1-Ceg1 interaction* is...
	A and *B* interact	*Stn1 and Cdc13* proteins displayed a physical *interaction* by...
Associate	*A* associate with *B*	*Atx1* also *associated* directly *with* the cytosolic domains of *Ccc2*.
	association between *A* and *B*	Physical *association between GCN5 and ADA2*.
	association of *A* (with|and) *B*	*Association of Vma12p* with *Vph1p*.
	A and *B* association with each other	The *SET4 and STE18* gene products *associated with each other*.
Bind	*A* bind to *B*	*GCN binds to ADA2*...
	bind of *A* to *B*	The *binding of Met28* to *DNA*.
	A and *B* bind	*Cdc24p and Bem1p bind* to each other
	bind between *A* and *B*	*Binding between TIF34* and *TIF35* in_vitro.
	A bind *B*	the N-terminal of *SINI* is suffisient to *bind SAP1*.
Complex	*A*(-|/)*B* complex	*Pc11, 2-Pho85* kinase *complexes* become essential...
	A and *B* complex	*Cdc46p* and *Cdc47p*... *complex* with each other.
	complex *A* and *B*	*Poll* and *Pob3* may form a *complex*...
	A complex with *B*	*GCG20* was... *complex* formation *with GCN1*.
	A complex... contain *B*	*Boilp* is part of a larger *complex* that *contains Cdc42p*.
	A complex *B*	*Ste11 complexed* to *Ste7*...

FIGURE 4.6 – Exemple de patrons linguistiques pour l'identification d'interaction entre gènes [OHTT01]

linguistique possible pour identifier une interaction ou une association... Il atteignent ainsi sur deux corpus des taux de rappel de plus de 80% et une précision de plus de 93%.

Les règles, dans le cadre de [OHTT01] et comme dans beaucoup d'approches, sont écrites à la main et affinées par essais-erreurs. Il s'agit donc d'une tâche longue et coûteuse. Il existe plusieurs environnements dédiés au développement de systèmes d'extraction d'information comme GATE (General Architecture for Text Engineering), OpenCalais ou Mallet (Machine Learning for Language Toolkit).

Une règle traduit généralement une relation entre deux entités et les problèmes plus complexes sont décomposés en plusieurs problèmes simples, dans le but d'obtenir des règles plus robustes. À l'inverse, des approches un peu plus ancrées dans la linguistique exploitent les structures prédicat-argument dans les phrases [YTMT01, PCZ+02] et les mettent en correspondance avec une structure conceptuelle. La figure 4.7 propose deux règles de mise en correspondance et la figure 4.8 extraite de [YTMT01] donne un exemple de mise en œuvre d'une règle.

Dans [YTMT01], une illustration de cette mise en correspondance entre structure linguistique et structure de frame est donnée par le schéma général du système repris en figure 4.8.

Contrairement à l'extraction d'entités nommées, il y a encore assez peu de travaux à exploiter des techniques d'apprentissage pour l'extraction de règles. Les travaux de Craven et Kumlien [CK99] que je décris ici, bien qu'anciens, montrent bien que disposer d'un éventail assez large de méthodes d'apprentissage ne suffit pas à extraire l'information pertinente des textes. La préparation des textes et la définition des marqueurs utilisés par ces algorithmes sont probablement encore plus importantes.

L'algorithme d'apprentissage utilisé dans [CK99] est un algorithme d'apprentissage relationnel inductif, similaire à FOIL [Qui90]. Les textes sont analysés phrase par phrase,

Target Verbs:

"bind", "make (complex with)"

Mapping Rules:

$$\begin{bmatrix} \text{REL} : \textit{"bind"} \\ \text{ARGS} : \begin{bmatrix} \text{SUBJ} : [1] \\ \text{COMPS} : [2] \end{bmatrix} \end{bmatrix} \rightarrow \begin{bmatrix} \text{FRAME} : \textit{"bind"} \\ \text{SLOTS} : \begin{bmatrix} \text{BINDER} : [1] \\ \text{BINDEE} : [2] \end{bmatrix} \end{bmatrix}$$

$$\begin{bmatrix} \text{REL} : \textit{"make"} \\ \text{ARGS} : \begin{bmatrix} \text{SUBJ} : [1] \\ \text{COMPS} : \textit{"complex"} \end{bmatrix} \\ \text{ADJ} : \textit{"with [2]"} \end{bmatrix} \rightarrow \begin{bmatrix} \text{FRAME} : \textit{"bind"} \\ \text{SLOTS} : \begin{bmatrix} \text{BINDER} : [1] \\ \text{BINDEE} : [2] \end{bmatrix} \end{bmatrix}$$

FIGURE 4.7 – Mise en correspondance entre des structures prédicat-argument [YTMT01]

tout d'abord en associant aux mots leur partie de discours puis en construisant l'arbre syntaxique de la phrase. L'analyseur Sundance [Ril98], illustré par la figure 4.9, est utilisé pour cette étape. [CK99] se propose de rechercher la localisation d'une protéine dans certains types de cellules ou dans certains tissus et les auteurs cherchent donc à définir un prédicat $\texttt{localisation} - \texttt{sentence}(\texttt{Sentence} - \texttt{ID}, \texttt{Phrase} - \texttt{ID}_1, \texttt{Phrase} - \texttt{ID}_2$ qui signifie que la phrase `Sentence-ID` permet de localiser $\texttt{Phrase} - \texttt{ID}_2$ dans $\texttt{Phrase} - \texttt{ID}_1$.

L'analyse syntaxique de la phrase (voir Fig 4.9) est utilisée pour instancier un ensemble de cinq types de prédicats qui seront donnés en entrée au processus d'apprentissage. La figure 4.10 donne les prédicats extraits de la phrase *"By immunofluorescence microscopy the PRP20 protein was located in the nucleus."*. Le choix de ces prédicats est crucial et conditionne le succès du processus. :

- $\texttt{phrase} - \texttt{type}(\texttt{Phrase} - \texttt{ID}, \texttt{Phrase} - \texttt{Type})$: associe un type à un syntagme (nominal, prépositionnel...) ;
- $\texttt{next} - \texttt{phrase}(\texttt{Phrase} - \texttt{ID}_1, \texttt{Phrase} - \texttt{ID}_2)$: reproduit l'ordre des syntagmes dans la phrase ($\texttt{Phrase} - \texttt{ID}_2$ succède à $\texttt{Phrase} - \texttt{ID}_1$) ;
- $\texttt{constituent} - \texttt{phrase}(\texttt{Phrase} - \texttt{ID}_1, \texttt{Phrase} - \texttt{ID}_2)$: établit que le syntagme $\texttt{Phrase} - \texttt{ID}_2$ est un sous-constituant du syntagme $\texttt{Phrase} - \texttt{ID}_1$;
- $\texttt{subject} - \texttt{verb}(\texttt{Phrase} - \texttt{ID}_1, \texttt{Phrase} - \texttt{ID}_2)$ et $\texttt{verb} - \texttt{direct} - \texttt{object}(\texttt{Phrase} - \texttt{ID}_1, \texttt{Phrase} - \texttt{ID}_2)$: permet de relier le sujet $\texttt{Phrase} - \texttt{ID}_1$ à son verbe $\texttt{Phrase} - \texttt{ID}_2$ ou le verbe $\texttt{Phrase} - \texttt{ID}_1$ à son complément d'objet direct $\texttt{Phrase} - \texttt{ID}_2$;
- $\texttt{same} - \texttt{clause}(\texttt{Phrase} - \texttt{ID}_1, \texttt{Phrase} - \texttt{ID}_2)$: relie des syntagmes apparaissant dans une même proposition.

De plus, un typage des mots est réalisé par 4 classifieurs bayésiens. Enfin, la figure 4.11 donne l'exemple d'une règle apprise par le système. Une telle règle est satisfaite si tous les prédicats en partie droite de la règle (situé à droite de " :-") sont satisfaits. Cette règle se traduit donc de la façon suivante :

- les deux premiers littéraux sélectionnent des phrases où le syntagme introduisant une protéine doit précéder le syntagme introduisant une localisation intra-cellulaire

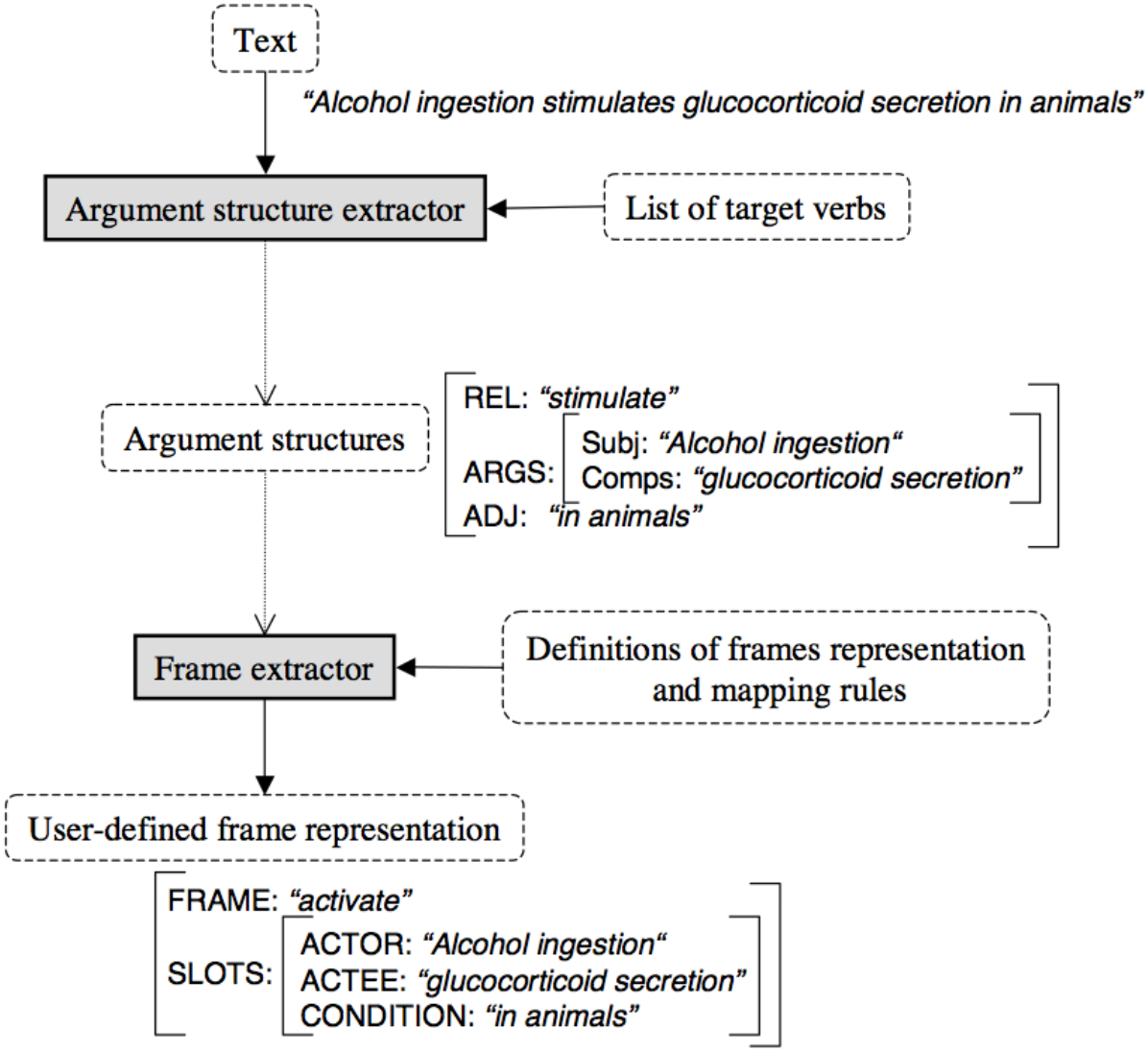


FIGURE 4.8 – Schéma global du système et exemple de transformation depuis la phrase juqu'à la structure de frame extraite [YTMT01]

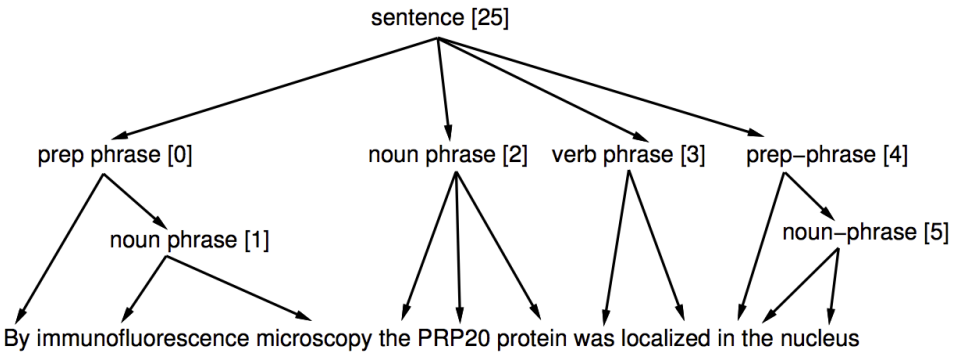


FIGURE 4.9 – Analyse syntaxique d'une phrase par Sundance [Ril98]

```
phrase-type(phrase-0, prepositional-phrase).
phrase-type(phrase-1, noun-phrase).
phrase-type(phrase-2, noun-phrase).
phrase-type(phrase-3, verb-phrase).
phrase-type(phrase-4, prepositional-phrase).
phrase-type(phrase-5, noun-phrase).

next-phrase(phrase-0, phrase-2).
next-phrase(phrase-2, phrase-3).
next-phrase(phrase-3, phrase-4).

constituent-phrase(phrase-0, phrase-1).
constituent-phrase(phrase-4, phrase-5).

subject-verb(phrase-2, phrase-3).

localization-sentence(sentence-25, phrase-2, phrase-5).
```

FIGURE 4.10 – Transformation de l'arbre syntaxique en prédicats en vue de la phase d'apprentissage ([CK99])

et ces deux syntagmes sont séparés par un autre syntagme (sans contrainte de type) ;
- le littéral suivant impose à la phrase d'être classé comme phrase par un classifieur bayésien ;
- le quatrième littéral doit reconnaître le syntagme désignant la protéine ;
- les deux derniers littéraux constituent les conditions pour une localisation intracellulaire.

Au final, il est intéressant de voir qu'en exploitant des littéraux de bas niveau, ne nécessitant pas une analyse linguistique fine, les résultats expérimentaux donnent une très bonne précision (de l'ordre de 92%) mais un faible rappel (21%).

```
localization-sentence(Sentence, Protein-Phrase, Location-Phrase) :-
                        next-phrase(Protein-Phrase, Phrase-1),
                        next-phrase(Phrase-1, Location-Phrase),
                        sentence-naive-bayes-1(Sentence),
                        phrase-naive-bayes-1(Protein-Phrase),
                        phrase-naive-bayes-2(Location-Phrase),
                        phrase-naive-bayes-3(Location-Phrase).
```

FIGURE 4.11 – Exemple de règle apprise par [CK99]

4.4 Confiance accordée à l'information

La notion de confiance accordée à l'information extraite d'un texte est une question intéressante abordée dans [Jil09] en recherchant les marqueurs de modalité épistémique. Nous en reprenons ici quelques éléments. Ces marqueurs appelés hedges dans [Lak72], ont fait l'objet d'une étude détaillée menée sur des articles en biologie cellulaire et moléculaire. Il identifie trois types de marqueurs :

- *X may cause Y* qui montre que l'information est le résultat d'un raisonnement et ne reflète pas nécessairement une connaissance certaine

– *These data indicate that...* qui montre que la conclusion est soumise à l'acceptation des données observées
– *we propose that...* qui suggère une vue alternative plutôt que définitive et valide.

Il semble que la plupart des informations extraites puissent être associées à un indice de confiance élevé. Cependant, il serait intéressant de confronter les indices de confiance accordés par différents auteurs à une même information. Ceci rejoint une autre question qui est de savoir comment traiter une information dans un article et son contraire, dans un autre article.

4.5 Ma contribution en extraction d'information

Mes travaux en extraction d'information se sont limités au développement d'un ensemble de règles sous GATE pour l'extraction d'information dans des textes de microbiologie et ne contribuent pas réellement à la recherche dans le domaine. C'est dans le cadre d'un CDD d'ingénieur de 6 mois que Bertrand Delecroix a développé ces travaux, repris, par la suite, dans le cadre de la thèse de Rokia Bendaoud.

Les textes sont des résumés d'articles scientifiques issus de la base Pascal de l'INIST et traitent de la mutation génétique des bactéries en résistance aux antibiotiques. L'identification des objets du domaine est faite dans GATE par l'utilisation de diverses nomenclatures. La figure 4.12 identifie dans un texte 9 types d'objets : les acides aminés, les antibiotiques, les bactéries, les codons, les enzymes, les gènes, les gènes mutants, les plasmides, et le protéines.

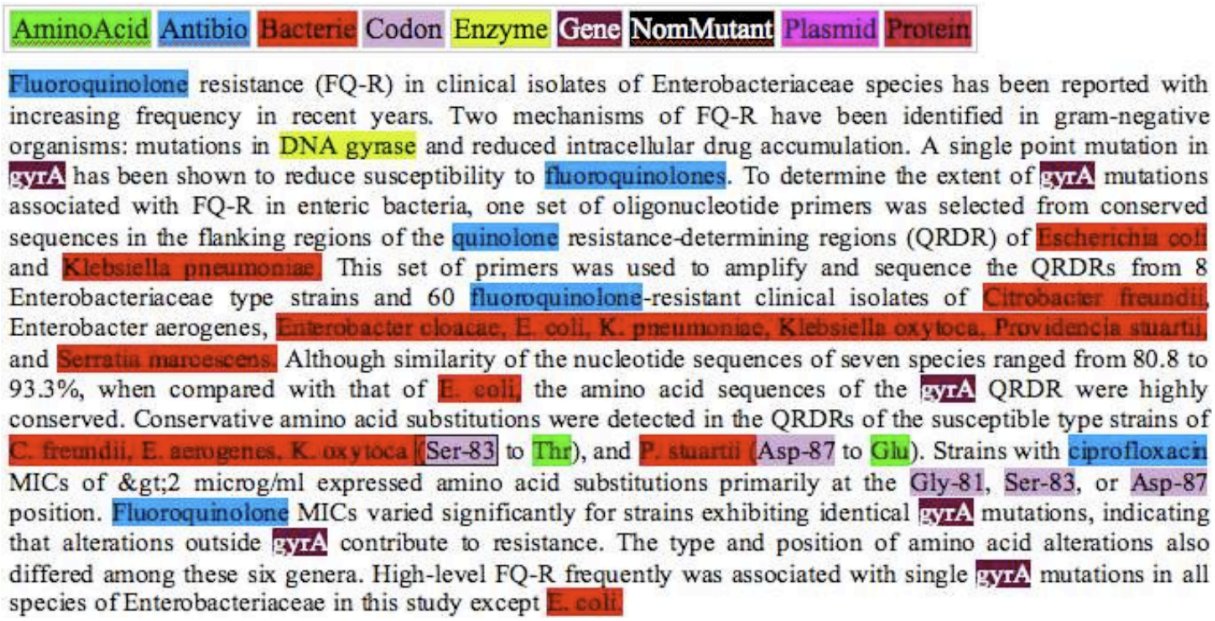


FIGURE 4.12 – Neuf types d'objets identifiés dans des textes de microbiologie

Nous nous sommes limités à l'extraction de deux relations : la relation de résistance d'une bactérie à un antibiotique et la relation entre un gène et une bactérie. 41 règles ont été définies pour identifier la première relation et 37 règles pour la seconde.

L'extraction d'information n'était pour nous qu'un moyen d'obtenir des données relationnelles pour les intégrer à un processus de construction d'ontologie reposant sur l'analyse relationnelle de concepts. N'ayant pas de corpus de référence et faute de temps, nous n'avons pas évalué cette étape selon les mesures de précision et de rappel habituelles.

4.6 conclusion

L'évaluation de l'extraction d'information doit être replacée dans le contexte de la construction d'ontologie à partir de textes. Dans la plupart des cas, il faut privilégier une bonne précision, au détriment d'un bon rappel.

Certains textes sont redondants (ou partiellement redondants) et les formes linguistiques au travers desquelles s'exprime cette redondance sont souvent différentes ce qui permet alors au système d'extraire malgré tout l'information pertinente. Mooney estime [NM00] que la différence entre les connaissances extraites, au final, est peu dépendante du choix manuel ou automatique pour la phase d'extraction d'information et que la part de bruit est comparable. Certains auteurs tentent aussi de réduire la dispersion et de pallier certaines formes de silence (ou d'informations manquantes) avant le processus de fouille. C'est le cas de [AR03] qui propose un processus de discrétisation pour passer d'un vecteur (TF/IDF) à un contexte formel, ou de [CHS05] qui introduit une étape de classification afin d'unifier la description des objets appartenant à une même classe. Ce lissage apporte, dans le cadre assez limité de leur expérimentation, de meilleurs résultats.

Au final, comme le souligne [NN05], l'extraction d'information et la construction d'ontologie sont très clairement reliées et doivent se concevoir dans un processus global itératif. l'IE devient donc un processus guidé par l'ontologie et la connaissance extraite est capitalisée dans l'ontologie. C'est une perspective que je propose de développer dans mon projet de recherche.

Chapitre 5

Projet de recherche

Plusieurs méthodes de fouille de données ont été présentées dans ce mémoire et je me suis intéressé exclusivement aux méthodes symboliques, à base de patrons. Mes travaux de recherche à venir visent en effet à explorer ces approches pour l'extraction de connaissances à partir de textes. Les expériences que j'ai menées ces dernières années, essentiellement sur des textes dans des domaines scientifiques, ont montré que les méthodes formelles permettent d'extraire une connaissance qui n'a pas toujours été identifiée auparavant par les experts, une connaissance relativement complexe, comme, par exemple, lorsqu'il s'agit d'expliquer certains phénomènes de mutation génétique des bactéries en résistance aux antibiotiques. Cependant, force est de constater que si nous étions capable de calculer un ratio entre les connaissances contenues dans un texte et celle que nous arrivons aujourd'hui à extraire puis à représenter, ce taux serait probablement très faible. Plusieurs raisons à cela : d'une part la difficulté à définir pour un ensemble de textes quelles sont les informations pertinentes puis, la difficulté à les extraire ; d'autre part, la complexité des structures encodant cette information et qui rend très coûteuse la fouille de données.

Dans la suite de cette section, je définis trois directions de recherche autour de la fouille de graphes, de la représentation de connaissances et de la définition d'un processus d'extraction de connaissances à partir de textes centrées sur l'analyse formelle de concepts. Je mentionne quelques sujets de thèse qui pourraient se mettre en place très rapidement. Je termine par la présentation de projets de recherche dans lesquels ces travaux vont s'intégrer.

5.1 Fouille de graphes pour la fouille de textes

La représentation d'un texte par un ensemble d'objets caractérisés par des propriétés binaires en vue de la construction du contexte formel exploitable par l'analyse formelle de concepts est très contraignante et extrêmement restrictive. L'analyse relationnelle de concepts, quant à elle, étend les possibilités de la FCA en prenant en compte des relations binaires. Chaque relation peut être définie à l'intérieur d'un même ensemble d'individus (le domaine et le co-domaine de la relation sont identiques) ou entre deux ensembles différents permettant alors de relier deux treillis décrivant deux domaines différents.

Les travaux récents sur les structures de patrons et sur la fouille de graphes ouvrent

de nouvelles perspectives. Les graphes ont été très souvent utilisés pour représenter des textes, que ce soit pour représenter leur structure (sous forme d'arbre XML, par exemple) ou pour en représenter le contenu, par exemple, à l'aide de réseaux sémantiques ou de graphes conceptuels. L'idée est donc de représenter le contenu d'un énoncé textuel sous la forme d'un graphe, de comparer ce graphe à d'autres graphes en recherchant, par exemple des sous-graphes communs, puis de classifier ces sous-graphes par l'analyse formelle de concepts en exploitant les bases posées par les travaux sur les structures de patrons.

5.1.1 Les structures de patrons (Pattern Structures)

Les treillis sont construits à partir d'une table binaire. Ainsi, dans une base de données décrivant des patients, l'âge d'un patient qui est un attribut numérique va devoir être transformé en un attribut binaire. Le plus souvent, l'âge exact du patient ne joue pas un rôle déterminant, par exemple, lorsque l'on s'intéresse aux effets indésirables des médicaments pour certaine classes de patients ; on va donc plutôt s'intéresser à des tranches d'âge. Il est alors possible d'appliquer un scaling conceptuel pour transformer en attributs binaires des attributs à valeurs numériques ou plus généralement en des attributs multivalués. Mais, le scaling, comme toute opération de discrétisation, s'accompagne d'une perte d'information comme le rappelle [Kay11].

Au lieu de travailler sur des données discrétisées, l'idée des structures de patrons est de travailler directement sur les données initiales et de définir des opérateurs de similarité. Ganter et Kuznetsov [GK01, Kuz09] donnent un cadre général dans lequel il suffit de définir un ordre partiel (demi-treillis supérieur) entre les attributs pour pouvoir construire le treillis de concepts. [Kay11] reprend ces travaux pour l'extraction de connaissances à partir de données biologiques.

Formellement, soit $\mathtt{G}$ l'ensemble d'objets, soit $(\mathtt{D}, \sqcap)$ un demi-treillis-supérieur décrivant les objets et soit $\delta : \mathtt{G} \longrightarrow \mathtt{D}$ la fonction associant chaque objet à sa description dans $\mathtt{D}$. $(\mathtt{G}, (\mathtt{D}, \sqcap), \delta)$ est appelé structure de patrons. Les éléments de $\mathtt{D}$ sont des patrons ordonnés par la relation de subsomption $(\sqsubseteq) : \forall \mathtt{c}, \mathtt{d} \in \mathtt{D},\ \mathtt{c} \sqsubseteq \mathtt{d} \iff \mathtt{c} \sqcap \mathtt{d} = \mathtt{c}$.

Les concepts de patrons de $(\mathtt{G}, (\mathtt{D}, \sqcap), \delta)$ sont des paires de la forme $(\mathtt{A}, \mathtt{d})$, $\mathtt{A} \subseteq \mathtt{G}$, $\mathtt{d} \in (\mathtt{D}, \sqcap)$, où $\mathtt{d}$ est la description commune à tous les objets de $\mathtt{A}$. L'ensemble des concepts de patrons forme un treillis.

Ces structures de patrons peuvent alors s'appliquer à des graphes, dès lors qu'un ordre partiel entre graphes est défini.

5.1.2 Fouille de graphes

Les graphes permettent de représenter des données complexes, de nature relationnelle ou multi-relationnelle comme par exemple en chimie, pour la recherche de sous-structures communes à des molécules ou la recherche de structures partagées par des réactions [PNV+10], ou en réseaux... La fouille de graphes, en raison de la complexité inhérente aux graphes (problème NP) fait l'objet de nombreux travaux [CF06, CH07, Bor09] soit pour mettre en place des structures de données performantes pour la recherche de sous-graphes [YH02], soit pour exploiter des heuristiques pour contraindre la recherche de sous-graphes [JCSZ10].

Ainsi, l'algorithme gspan [YH02] sert souvent de base à d'autres algorithmes. Il prend en entrée un ensemble de graphes connectés, étiquetés et non-orientés. De façon similaire à la recherche de motifs fréquents, c'est-à-dire des motifs partagés par un seuil minimum d'objets de la base de données, un sous-graphe est fréquent si le nombre de graphes le contenant est au dessus d'un certain seuil fixé par l'utilisateur. Étant donné une base de données de graphes étiquetés et un support minimum σ_s un sous-graphe sg est fréquent dans LG si et seulement si le support de sg est supérieur à σ_s. La recherche de sous-graphes repose sur la notion d'isomorphisme de graphes. L'optimisation de l'algorithme exploite la représentation des graphes sous la forme d'un arbre DFS (Deep-First-Search) choisi de façon à ce qu'un graphe n'ait qu'un seul DSF Code. Pour réduire l'espace de recherche, seuls les sous-graphes connectés sont recherchés.

Dans notre contexte, les graphes sont utilisés pour représenter des textes. Il nous faut donc accorder les possibilités des algorithmes de recherche de sous-graphes avec la représentation d'un texte sous forme de graphes.

5.1.3 Représentation d'un texte sous forme de graphes

Les graphes sont assez largement utilisés pour représenter des documents. D'un point de vue structurel, une structure XML est en elle-même un arbre. En extraction d'information, les formulaires (frames) peuvent être eux aussi considérés comme des arbres ou plus généralement comme des graphes si l'on considère que chacun des éléments de ces formulaires sont également des éléments décrits dans une ontologie. En représentation de connaissances et en TAL, les réseaux sémantiques ou les graphes conceptuels [Sow84, Sow08] proposent de représenter sous forme de graphes, des expressions de la logique du premier ordre.

Dans le cadre de ce projet et dans un premier temps, la notion de graphe doit donc être précisée et délimitée. Par exemple, [JCSZ10] propose une représentation des textes combinant différentes dimensions (représentation d'aspects syntaxiques et de l'ordre des mots dans la phrase, prise en compte d'étiquettes sémantiques associées aux mots...). De plus, les élements du graphe sont pondérés ; par exemple, les mots sont associés à leur valeur TF.IDF. Les auteurs ont ensuite défini la notion de support pondéré combinant le support d'un graphe et son poids, notion qu'ils ont ensuite intégrée à l'algorithme gspan.

D'autres travaux sur la catégorisation de documents cherchent à capturer des informations cachées dans les documents en exploitant les informations de nature structurelle données par les étiquettes HTML [SBLK05, MLK07, MLK08]. Lorsque ces informations sont de nature structurelle, linguistique ou sémantique, les termes du document sont introduits comme nœud dans le graphe. C'est vers ce type de représentation que nous nous orientons.

En effet, dans notre perspective d'extraction de connaissances à partir d'un texte, une première étape pourrait être de représenter un texte comme un réseau de concepts reliés par des relations, relations de nature sémantique ou ontologique, puis d'envisager progressivement la construction d'une représentation plus riche en fonction des besoins. Une première expérimentation a été réalisée dans le cadre d'un stage de première année de Master sur des recettes de cuisine (voir section 5.4.1). Un autre cadre d'expérimentation nous est donné par le projet Hybride sur les maladies rares (voir section 5.4.2).

En revanche, je pense en effet que la fouille de graphes telle que nous l'avons présentée n'est pas un cadre approprié à la fouille de graphes conceptuels. D'une part, les graphes conceptuels ne peuvent pas se réduire à des graphes connectés et étiquetés et d'autre part, il existe des travaux sur la FCA appliquée à des formules logiques [Fer02, FF10] qui seraient probablement plus appropriés à cet objectif.

Sujet de thèse : Fouille de graphes pour l'extraction de connaissances à partir de textes. La découverte de connaissances se heurte à un problème récurrent : le très grand nombre de motifs ou de règles d'association extraits par les algorithmes d'extraction de motifs. Cela est d'autant plus vrai lorsque les données sont des textes ou des graphes. Il est alors quasi-impossible pour un expert d'analyser les résultats du processus de fouille. Il existe cependant plusieurs méthodes pour diminuer ce nombre. L'une d'entre elles est d'augmenter le seuil de fréquence pour l'extraction de motifs fréquents mais les motifs très fréquents ne sont pas nécessairement les plus intéressants et les motifs un peu moins fréquents ne seront plus extraits. Une autre approche consiste à spécifier des contraintes ou à utiliser des mesures pour filtrer les motifs extraits. Parmi ces mesures, nous reprendrons puis affinerons la notion de "motifs les plus informatifs" proposée dans le cadre de la fouille de structures moléculaires [PNV+10].

Les motifs intéressants sont les motifs qui sont discriminants c'est-à-dire (1) des motifs capables de représenter une famille importante de motifs (par rapport à la factorisation ou à la cohésion interne) et (2) des motifs distincts des autres familles (discrimination). La notion de motifs les plus informatifs répond bien au besoin de discrimination. La recherche des motifs les plus informatifs peut donc être vue comme un compromis entre la fréquence des motifs et leur valeur informative. Une mesure de ce compromis peut être réalisée par une fonction de score. Dans le cas d'une collection de textes, le problème est de chercher les motifs de graphes ayant un score élevé.

La notion de motifs les plus informatifs a été introduite pour guider la fouille de sous-graphes. L'idée ici est d'étendre ces travaux à l'analyse de textes et au traitement automatique de la langue en s'appuyant sur des techniques de fouille de graphes. Les graphes sont en effet bien appropriés à la représentation d'un texte que ce soit pour rendre compte d'aspects structurels ou sémantiques [JCSZ10]. Il faut donc définir une fonction de score qui prend en compte des connaissances du domaine et des connaissances de nature linguistique.

À partir d'un ensemble de motifs de sous-graphes extraits d'un corpus de textes en utilisant les motifs les plus informatifs, l'idée est d'étendre la notion de structure de motifs en FCA à des sous-graphes. La classification de textes relativement aux graphes qui les décrivent sera ainsi organisée en un treillis. Différentes applications peuvent alors être envisagées, comme la recherche d'information ou encore la construction d'ontologie. Cette classification suppose cependant d'adapter les travaux réalisés jusqu'à présent sur les structures de motifs [Kuz09, Kay11].

Cette thèse va donc établir un pont entre deux approches de fouille, la fouille de sous-graphes d'une part, et l'analyse formelle de concepts d'autre part et les appliquer à l'extraction de connaissances à partir de textes. Ce sont deux approches très puissantes, leur association permettra l'enrichissement de chacune d'entre elles et au final, les connais-

sances extraites seront plus complètes et précises.

5.2 La représentation de connaissances

Le formalisme de représentation de connaissances que j'ai retenu est les logiques de descriptions et, je ne les étudierai pas en soi mais pour leurs liens avec l'analyse formelle de concepts. Les logiques de descriptions offrent à la fois un cadre de représentation formel et des outils de raisonnement performants. Rouane-Hacène [HNRV07a, RH08], repris en partie par Bendaoud [Ben09], propose une transformation de l'analyse relationnelle de concepts vers une logique de descriptions $\mathcal{FL}$. L'idée est de proposer une transformation des entités manipulées en FCA/RCA (individus, propriétés, relations, concepts du treillis et relations entre concepts) en un ensemble de concepts atomiques ou définis et de rôles. L'exemple d'Oedipe [BCM+02] (p.73) illustre une des différences entre FCA et LD sur les déductions que l'on peut faire dans un monde fermé (approche "base de données") versus dans un monde ouvert (approche logique). De façon plus générale, [Ser08] dresse un panorama assez détaillé des liens entre FCA et LD que je souhaite approfondir. En effet, alors que dans les travaux que j'ai menés jusqu'à présent, les connaissances extraites étaient essentiellement interrogées par un langage de type SPARQL, dans le cadre du projet KolFlow (voir section 5.4.3), les connaissances extraites des textes se trouvent au cœur d'un processus de raisonnement, *i.e.* un processus de raisonnement à partir de cas exploitant une ontologie pour adapter des recettes de cuisine. Il s'agit donc d'un cadre propice à une expression des connaissances plus riche.

Par ailleurs, comme je le soulignais à propos des graphes conceptuels, un certain nombre de travaux se sont intéressés aux liens entre l'analyse formelle de concepts et la logique. Ainsi, Chaudron *et al.* [CM98] ou Ferré [Fer02] prennent en compte des objets décrits par des propriétés exprimées sous forme de formules logiques. La classification de formules logiques présente clairement un intérêt pour se rapprocher du TAL, d'énoncés en langue naturelle représentés sous une forme logique [FF10]. Ces questions restent cependant un projet à moyen voire long terme.

5.3 La FCA au cœur d'un processus continu, itératif et interactif

L'utilisation de connaissances dans les systèmes informatisés n'est plus réservée aux applications en intelligence artificielle et se généralise, on trouve ainsi des ontologies depuis les couches *"basses"* de systèmes comme le niveau réseau jusqu'aux applications de plus *"haut"* niveau comme la composition de services web ou la gestion des dossiers patients dans un hôpital. Alors que les sources de données, et notamment, les sources textuelles continuent de croître de façon exponentielle, la production de connaissances reste un processus complexe et coûteux. De plus, les connaissances produites ne dépendent pas que des ressources mais également de la tâche pour laquelle elles sont utilisées. Cela signifie donc que dans le processus de construction, il faut pouvoir infléchir le processus vers tel ou tel "aspect". C'est souvent le rôle de l'utilisateur ou de l'expert. Il reste donc beaucoup

à faire pour que l'acquisition, l'utilisation et la mise à jour continue des connaissances se fasse en temps quasi-réel et de façon continue pour répondre aux besoins d'évaluation des experts et prévenir l'obsolescence des connaissances.

Nous avons montré dans les chapitres précédents que la fouille de textes se divise en deux grandes étapes. L'extraction d'information transforme le texte en des données généralement de nature symbolique dans lesquelles on identifie les objets ou individus, leurs propriétés (ou attributs), et les relations (souvent binaires) entre ces individus. L'extraction de connaissances exploite alors, grâce à des méthodes de fouille de données, ces informations pour construire des modèles, qui, une fois interprétés et validés par des experts deviendront des connaissances. Il nous faut cependant introduire une troisième étape, l'annotation sémantique qui exploite ces connaissances pour repérer dans les textes les concepts et permet alors une exploitation conceptuelle des textes.

Les méthodes sur lesquelles ces différentes étapes reposent, comme nous l'avons vu dans les chapitres précédents, ont des fondements théoriques très différents (réseaux bayésiens, extraction de motifs...). Dans un processus que l'on souhaite itératif et où un expert ou un utilisateur souhaite progresser au cours des itérations vers une analyse de son domaine de plus en plus fine (en espérant soit atteindre un point fixe, soit éviter la régression du système), cette diversité des méthodes entre les étapes est difficile à gérer, notamment parce qu'il est difficile de connaître l'impact sur une étape des modifications qui ont été effectuées sur une autre étape. Ainsi, il peut apparaître qu'une propriété extraite au cours de l'extraction d'information ne devrait plus être extraite, problème assez similaire à l'exemple de la figure 3.4 dans laquelle il fallait fusionner les deux propriétés `Neutral Gram` et `Positive Gram`. De tels changements peuvent, par exemple, nécessiter un nouvel entraînement des outils d'extraction d'information.

L'utilisation de méthodes symboliques et reposant sur les mêmes fondements d'une étape à l'autre prend alors tout son sens. Les premières étapes reposent, comme nous l'avons vu sur les mots ou les termes qui permettent de construire un premier niveau de conceptualisation qui peut alors être utilisé pour annoter des textes. L'annotation des textes par les concepts peut alors devenir le niveau d'information à prendre en compte pour effectuer de nouvelles opérations de fouille de textes.

L'intérêt d'une telle approche, est double :

- un cadre unifié et formel pour l'extraction de connaissances assure une construction des connaissances cohérente (du point de vue logique). De plus, la représentation des connaissances par un formalisme logique permet à la fois à la machine et à l'humain de raisonner.
- Le continuum dans les différentes étapes du traitement des connaissances assure une traçabilité entre ces étapes. Il permet de concevoir la fouille de textes comme une construction itérative et interactive des connaissances et des annotations des textes.

Comme nous le soulignons en section 3.4.1et dans [BNT08b], des informations de natures différentes peuvent contribuer à l'extraction de connaissances : relations hiérarchiques entre concepts identifiées par des experts, des thésaurus ou des ontologies, propriétés caractérisant des objets, relation entre objets, voire même graphes caractérisant des objets. Les sources contenant ces informations peuvent être des textes ou des fragments d'ontologies déjà existantes venant compléter les informations textuelles.

L'analyse formelle de concepts est un outil pour la conceptualisation d'un domaine

particulièrement puissant que je souhaite placer au cœur du processus d'extraction de connaissances et qui peut prendre en compte ces différents types de ressources et de connaissances. Le treillis de concepts peut ensuite être transformé en une base de connaissances écrite en logique de descriptions dans laquelle certains concepts sont exprimés sous forme de conditions nécessaires et suffisantes.

La Figure 1.1 introduite en page 10 présente l'extraction de connaissances comme un processus itératif et interactif. En réalité, le processus d'extraction de connaissances est de fait relié à d'autres processus. La figure 5.1 considère le processus d'extraction de connaissances comme le pendant de l'annotation sémantique. En effet, l'annotation des textes et, en particulier, l'annotation sémantique est exploitée par l'extraction de connaissances notamment pour la construction d'un contexte formel. En retour, les nouvelles connaissances extraites sont exploitées par le processus d'annotation pour annoter de façon plus fine les textes. A cette boucle, il faut ajouter les contraintes issues de l'application pour laquelle l'ontologie est construite, recherche d'information "intelligente" ou raisonnement qui, par essai-erreur, vont amener l'expert à modifier l'ontologie. Utiliser la FCA pour la conceptualisation des connaissances, c'est s'assurer d'une certaine traçabilité entre les informations extraites des textes et les concepts de l'ontologie. C'est un objectif que nous nous sommes donnés dans le cadre du projet ANR Kolflow dans lequel je suis responsable de tâche 2 sur l'extraction continue de connaissances et l'annotation (voir section 5.4.3).

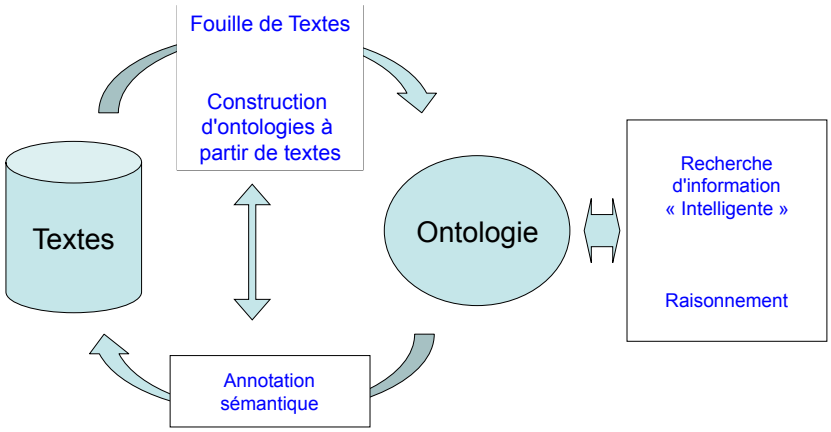


FIGURE 5.1 – Schéma général d'un processus de fouille de textes

Le défi qui se pose ici est donc de faire coopérer la machine et l'humain autour d'un système formel qu'est la FCA pour extraire et représenter des connaissances. On doit donc pouvoir à tout moment conceptualiser un domaine à partir d'informations extraites de corpus ou d'autres sources de connaissances. A l'inverse, un utilisateur – un expert – doit

pouvoir corriger et enrichir la conceptualisation proposée par la FCA. Cela se démarque des approches usuelles dans lesquelles, aux questions de cohérence près, un expert intervient directement et assez librement sur l'ontologie en utilisant un éditeur comme, par exemple, PROTÉGÉ. En contrepartie, les contraintes imposées par l'utilisation d'un système formel peuvent entacher sérieusement la dynamique d'extraction de connaissances. Les grandes questions que je souhaite donc aborder pour la réalisation d'un tel système sont les suivantes :

1. définir la FCA comme processus continu de conceptualisation ;
2. articuler extraction d'information et annotation sémantique autour de la FCA ;
3. augmenter la capacité des treillis à exprimer une connaissance "complexe" : la fouille de graphes et la classification de graphes en FCA présentée en section 5.1 doit apporter une première réponse à cette question.

5.3.1 La FCA comme processus continu de conceptualisation

L'idée maîtresse ici est de garder la structure formelle de treillis issue de la FCA comme élément central du processus d'extraction de connaissances. Cela suppose donc d'une part, de préserver à tout moment, le lien entre le treillis et l'ontologie représentée dans un langage de représentation des connaissances et d'autre part de préserver le lien entre les informations/annotations en entrée de la FCA et la conceptualisation souhaitée. Du point de vue de la dimension cognitive du processus de construction d'ontologie, c'est un cadre volontairement restrictif mais qui devrait permettre la définition de nouvelles opérations et de nouveaux outils pour la construction et l'enrichissement d'ontologies s'appuyant sur des bases formelles. L'utilisateur ou l'expert du domaine est alors sollicité pour évaluer la pertinence des concepts du treillis avant qu'ils ne soient traduits en logique de descriptions.

La construction d'ontologie à partir de textes a fait l'objet de très nombreux travaux. Parmi les plus proches de nos travaux, il faut citer les travaux de Maedche et Staab [MS00b, MS01b, MS01a] pour lesquels la FCA est utilisée comme méthode de conceptualisation. Placé au cœur d'un processus continu, il faut s'intéresser à la modification et à l'évolution des ontologies. L'évolution des ontologies peut se voir sous différents angles et depuis 2004, plusieurs travaux ont classifié les différentes évolutions possibles [Sto04, FPA06, Kle04]. Ainsi, dans le cadre du web sémantique, la dimension distribuée des ontologies posent des problèmes que [Kle04] divise en deux grandes classes : les problèmes liés au langage dans lequel l'ontologie est représentée ou les problèmes liés à l'organisation de l'ontologie avec notamment des différences liés à la conceptualisation ("concept scope" et "model coverage") ou à l'"explication" (paradigme pour la description des concepts, définition des concepts et notamment identification des concepts atomiques...). [FPA06] propose un état de l'art très complet sur l'évolution des ontologies et sur les raisons qui motivent cette évolution.

Dans ce projet, toutes ces dimensions ne sont pas prises en compte, l'objectif étant d'accorder les différentes étapes liées à l'extraction de connaissances et à rendre traçable ce processus. Quelques éléments de réflexion nous permettent d'étayer ce projet.

Définition d'opérations pour la modification du contexte formel

Il ne peut être construit qu'un seul et unique treillis à partir d'un contexte formel. Aussi, si l'on veut modifier le treillis (fusionner des concepts, séparer un concept en deux concepts distincts...), il faut en modifier le contexte formel. La difficulté pour l'expert en charge de ces modifications est de savoir comment modifier le contexte formel. Bien que sous une forme encore très simpliste, une première ébauche de ce travail a été proposée dans [BTN10] (voir section 3.4.4). L'expert est sollicité pour évaluer la pertinence des concepts, tant du point de vue de son extension que de son intension. Il lui est alors proposé de corriger le contexte formel pour que le treillis résultat s'accorde mieux avec ses attentes en créant ou en supprimant des propriétés associées à un seul objet ou à un ensemble d'objets. Le but est maintenant de pouvoir définir un ensemble d'opérations d'un peu plus "haut niveau", répondant à des besoins utilisateurs du type (la liste n'est pas exhaustive) :

- L'objet x_i ne devrait pas figurer dans l'extension du concept C_j ;
- L'attribut a_i devrait être partagé par tous les objets du concept C_j ;
- Le concept C_i et C_j devraient ne constituer qu'un seul et même concept.

Ces questions se traduisent en une ou plusieurs opérations élémentaires sur le contexte formel telles qu'introduites dans [BTN10], opérations qui nécessiteront parfois des interactions avec l'expert pour qu'il fournisse des informations additionnelles si nécessaires.

Incrémentalité dans la construction de treillis

Du point de vue de l'analyse formelle de concepts, plusieurs travaux se sont intéressés à la construction incrémentale des treillis. Ces algorithmes répondent aux situations où de nouvelles propriétés ou de nouveaux objets sont ajoutés au contexte formel. Nous avons évoqué en section 3.1.1 l'algorithme incrémental de Godin [GM94] qui évite que le treillis ne soit recalculé dans son intégralité ; l'algorithme calcule les modifications à apporter au treillis initial. [VM01] propose une généralisation des algorithmes incrémentaux pour la construction de treillis.

Si ces algorithmes peuvent être le point de départ de notre réflexion pour placer la FCA dans un contexte d'enrichissement continu, le problème à résoudre est bien différent et plus complexe, notamment parce que la relation d'incidence $\mathcal{I}$ peut être modifiée pour des attributs ou des objets déjà existants.

Sujet de thèse : La thèse vise à définir un processus où l'humain et la machine sont amenés à coopérer pour extraire des connaissances à partir de différentes ressources. L'enjeu est qu'à tout moment, les méthodes automatiques (et formelles) d'extraction de connaissances puissent être appelées et que les connaissances ainsi extraites puissent être mises en accord avec des connaissances déjà extraites. L'humain doit donc pouvoir corriger ou enrichir l'ontologie à tout moment. Dans la plupart des approches, lorsque la phase de conceptualisation est réalisée par un outil automatique, l'humain intervient après cette phase et adapte la conceptualisation à sa perception du domaine en utilisant un éditeur d'ontologie. Il a alors assez peu de contraintes formelles si ce n'est celle de respecter la cohérence logique de la base de connaissances.

L'extraction de connaissances est un processus itératif et interactif composé de plusieurs étapes. L'interaction est souvent associée à l'évaluation où l'on demande aux experts du domaine d'interpréter et de valider les patrons extraits par la phase de fouille de données. En réalité, l'extraction de connaissances se fait par essai-erreur et la correction des erreurs supposent généralement d'observer et de modifier les étapes antérieures à la fouille de données, comme la sélection du corpus ou la préparation des données. Nous souhaitons concevoir ces modifications comme une modification de l'annotation des textes. Ainsi, l'annotation sémantique guide la construction d'ontologie et inversement, l'ontologie guide l'annotation sémantique.

L'approche développée dans cette thèse exploite des méthodes formelles pour l'extraction de connaissances. Elle doit définir un processus continu d'extraction de connaissances reposant sur les travaux développés dans l'équipe Orpailleur. En effet, la thèse de Rokia Bendaoud exploite l'analyse formelle de concepts pour construire une conceptualisation du domaine à partir de textes. Après validation par les experts cette conceptualisation est transformée en une ontologie codée en logique de descriptions. Un prototype met en œuvre cette méthode et a été expérimenté en astronomie, en microbiologie et sur le domaine de la cuisine dans le contexte du projet Taaable. Le point fort d'une telle approche formelle est que le processus de conceptualisation guidé par les ressources du corpus est moins coûteux en temps et moins subjectif. Cependant, un tel processus souffre également de points faibles. Nous nous concentrerons sur trois d'entre eux :

- Comment les approches formelles comme la FCA peuvent prendre en compte l'interaction avec les experts ? En effet, un expert peut interagir assez librement avec un wiki ou un wiki sémantique mais cette interaction n'est pas nécessairement compatible avec la structuration formelle proposée. Pourtant, la structure formelle doit être maintenue à jour puisque c'est elle qui garantit que de nouveaux textes, de nouvelles ressources ou de nouvelles interactions peuvent être prises en compte à tout moment. C'est aussi au travers de cette structure formelle que peut se faire le lien avec l'annotation sémantique.
- L'annotation sémantique est à la base des wikis sémantiques. Dans un système continu d'extraction de connaissances, l'annotation est utilisée pour l'extraction de connaissance, notamment pour la construction d'un contexte formel. En retour, les connaissances extraites doivent permettre d'annoter plus finement les documents. Au cours des itérations successives, certaines annotations doivent être supprimées et d'autres mises à jour pour permettre l'évolution de l'ontologie.
- Le prototype pour la conceptualisation doit pouvoir s'insérer dans une chaine de traitement construite autour d'un wiki sémantique. Les textes sont donc des ressources de ce wiki. Les processus de traitement automatique de la langue ainsi que d'analyse formelle de concepts seront alors définis comme des modules. De même, un processus devra transformer la conceptualisation proposée par la FCA en une ontologie en prenant en compte les contraintes, notamment en terme d'expressivité, imposées par cet environnement.

La thèse se déroulera dans le contexte du projet Kolflow, un projet ANR de 3 ans et demi qui vient de débuter.

5.3.2 Articuler extraction d'information et annotation sémantique autour de la FCA

Pour préserver la traçabilité dans le processus d'extraction de connaissances, la modification du contexte formel doit s'appuyer sur une ou plusieurs sources d'information ou de connaissances qui permettront de générer le nouveau contexte formel associé au nouveau treillis. S'il s'agit d'une source textuelle, le processus d'extraction d'information doit être enrichi pour que cette nouvelle information puisse être identifiée dans les textes. S'il s'agit d'une source de connaissances externe, elle viendra compléter les connaissances déjà prises en compte. De façon similaire, le modèle de connaissances validé, l'annotation sémantique des documents textuels sera enrichie, contribuant ainsi à la continuité du processus.

L'extraction de motifs pour l'extraction d'information

La fouille de textes repose généralement sur la définition *a priori* d'un modèle de connaissances, modèle qui identifie les objets, les propriétés et les relations. L'extraction d'information intervient alors pour instancier le modèle dont les résultats sont utilisés pour l'extraction de connaissances, *i.e.* le processus de classification construisant les concepts à partir des informations. Mais, comme le souligne [NN05], la construction de l'ontologie et l'extraction d'information sont très liées et il y a nécessairement de nombreuses itérations entre ces deux étapes.

Nous porterons un intérêt tout particulier aux travaux actuels initiés par le GREYC autour des méthodes à base d'extraction de motifs et motifs séquentiels pour l'extraction d'information, un processus de nature symbolique permettant probablement une meilleure paramétrisation, c'est-à-dire, une meilleure sélection des objets et relations à extraire des textes. Ces travaux [PC09, PCK$^+$09] (voir section 2.5.3) reposent sur des méthodes d'extraction de motifs, de règles d'association et de treillis et relient ainsi extraction d'information et fouille de données. Ces méthodes peuvent être appliquées aux mots ou aux mots annotés par des informations de nature linguistique.

Sujet de thèse : Extraction d'information guidée par des connaissances Nous avons vu dans le chapitre 4.3.1 que l'extraction d'information cherche à identifier l'information dans les textes d'un domaine en apprenant puis en recherchant des indices linguistiques reflétant des attributs ou des relations. Cette thèse vise à proposer une méthode d'apprentissage à base d'extraction de motifs prenant en compte les connaissances déjà acquises dans des passes précédentes de fouille de textes.

Le parallèle peut être fait avec un système comme GATE qui extrait l'information des textes par application de règles dont la prémisse correspond à un patron ou des marqueurs qui doivent être identifiés dans les textes et la conclusion est une annotation d'une portion de texte par une certaine étiquette. Les règles appartiennent à des niveaux, de telle sorte qu'une règle à un niveau peut utiliser les annotations introduites par les règles de niveau inférieur (qui sont donc appliquées avant). GATE donne de bons résultats mais le temps d'écriture des règles, à la main, est long. De plus, ces règles ne peuvent refléter que ce que les experts recherchent, et donc, connaissent partiellement.

La plupart des méthodes d'apprentissages en extraction d'information reposent le plus souvent sur des approches numériques, notamment sur l'utilisation des SVN. Ces méthodes privilégient le rappel à la précision. Or, dans notre cas où l'extraction d'information est exploitée pour construire un contexte formel, la précision est une dimension essentielle. Tout élément annoté à tort introduit du bruit dans la classification par FCA. Cette thèse a donc pour point de départ les travaux de [PCK+09] qui extraient des textes les motifs séquentiels fréquents et des règles d'association. L'idée est de s'inscrire dans une démarche itérative du processus de fouille de textes dans lequel chaque itération permet d'enrichir une ontologie qui permet alors d'annoter les textes par des concepts. La nouvelle passe d'extraction d'information doit donc exploiter les concepts annotés dans les textes pour identifier de nouveaux motifs.

L'annotation sémantique

L'annotation sémantique consiste à identifier dans les textes les concepts du domaine [UCI+06]. Dans le contexte du web sémantique, de nombreux outils ont été proposés. Dans les synthèses proposées par [Ama07] ou [May07] certains systèmes sont spécifiques aux textes, d'autres à différents médias comme les images, les films...

L'annotation sémantique est perçue comme un processus prenant en entrée un texte, en ressource, une ontologie et produisant un texte annoté, c'est-à-dire, un texte décoré de métadata. Par rapport au continuum évoqué précédemment, l'annotation peut être vue comme la fin d'une itération dans le processus de fouille de textes et constituer alors, les données en entrée à l'étape suivante. C'est dans cette optique que je propose de placer l'annotation sémantique au sein du processus de fouille de textes. La méthode de construction des ontologies à partir de la FCA permet d'envisager l'annotation sémantique sous un angle particulier.

La FCA propose non seulement une conceptualisation du domaine mais également, comme nous l'avons vu dans la transformation des concepts en logique de descriptions, une définition des concepts, c'est-à-dire, des conditions nécessaires et suffisantes. Dans un certain nombre de situations, exploiter ces conditions nécessaires et suffisantes peut s'avérer intéressant. Par exemple, dans le cas de recettes de cuisine, on peut s'intéresser à définir ce qu'est une soupe (même s'il est probable qu'il existe plusieurs types de soupes différents), une sauce ou une marinade, au delà du fait que les mots soupe, sauce ou marinade apparaissent dans la recette. Dans un contexte médical, un ensemble de symptômes pourraient s'avérer être suffisant pour l'identification d'une maladie ou, au contraire, ne pas être suffisamment discriminant et ne permettre que l'identification d'un ensemble de maladies.

Sujet de thèse : Construction de prototypes ou de patrons d'ontologie Dans de nombreuses situations, il est plus important d'extraire des patrons ou des sortes de prototypes à partir d'un domaine que de proposer une ontologie complète. Les travaux actuels sur les "Ontology Design Patterns" [DMBSV02] visent à identifier des éléments de connaissances qui peuvent être utilisés (et réutilisés) pour la construction d'ontologies. Une question dans un tel contexte serait : est-il possible de définir ce qu'est une soupe à partir d'un ensemble de recettes de cuisine. Dans son stage de master 2ème année, Israel

Wakwoya [WB10] exploite la FCA et l'extraction de règles d'association pour identifier un tel patron d'ontologie. Dans la même lignée mais plus orienté vers les textes, Valmi Dufour [DLLNT10], dans un stage de M1 que j'ai encadré, s'est intéressé à modifier une recette de cuisine en exploitant un ensemble d'autres recettes de cuisine pour répondre à une question du type : si je remplace tel ingrédient par tel autre ingrédient, quelles actions dois-je supprimer et quelle préparation dois-je faire subir au nouvel ingrédient ?

Ces deux travaux montrent que la FCA est particulièrement bien adaptée à construire des généralisations et qu'elles peuvent être exploitées soit pour construire des patrons d'ontologies, soit pour les utiliser directement dans la transformation ou dans l'annotation de textes.

5.4 Ouvertures vers des domaines d'application

Cette section présente des projets dans lesquels je suis impliqué et qui fournissent des cadres d'expérimentation. Au travers de ces cadres d'expérimentation, ma volonté est de ne pas dédier mes travaux à un domaine en particulier mais de concevoir des outils génériques qui peuvent par la suite être configurés ou modifiés pour s'adapter à des domaines spécifiques. Bien évidemment, cette limite entre le dédié à un domaine et le générique est parfois un peu artificielle et difficile à définir précisément.

5.4.1 Le projet Taaable

Même si le domaine de la cuisine ne semble pas à première vue représenter un enjeu stratégique, ce domaine est à lui seul un condensé des problèmes que l'on rencontre dans pratiquement tous les autres domaines. Les textes de cuisine sont accessibles à tous et nous avons tous un bagage minimun permettant de comprendre rapidement les défis mais aussi les limites de nos travaux. Il y a de très nombreuses ressources textuelles, accessibles sur le web, qu'il s'agisse de la centaine de milliers de recettes, de thésaurus ou de pseudo-ontologies plus ou moins formalisées.

Le *Computer Cooking Contest* (CCC) est un workshop satellite de l'International Conference on Case-Based Reasoning (ICCBR). C'est un défi pour lequel un système opérationnel doit être réalisé en réponse à 3 tâches différentes :

- Les figures imposées : il faut être capable de retrouver et d'adapter si nécessaire une recette pour répondre à une requête. Une base de recettes "de référence" est imposée (en anglais). Les exemples de requêtes données sont *Préparer un plat principal avec de la dinde, des pistaches et des pâtes sans ail* et une réponse possible est de remplacer le poulet par de la dinde dans la recette *Pistachio Chicken* de la base.
- Le défi "adaptation" : il suppose d'adapter à la fois des ingrédients et les instructions de préparation pour un plat.
- Le défi "menu" : il s'agit de composer un menu de trois plats à partir de la base des recettes. L'exemple de question proposé est le suivant : *J'ai du filet de bœuf, des carottes et du céleri, de l'ail et des concombres. J'ai des pommes de terre également. Pour le dessert, il y a des oranges et de la menthe. Une soupe serait une bonne entrée.* Dans ce cas, trois recettes de la base pourraient être suggérées : un "Caldo

Verde" en entrée, un "filet steak with baked potatoes" en plat principal, et "an orange ice cream with mint flavour" en dessert.

Pour participer au concours CCC, le projet Taaable a associé la première année l'équipe Orpailleur du LORIA, une équipe du LIRIS à Lyon et une équipe du LIM&BIO à Paris, puis, les années suivantes, essentiellement le LORIA.

Le défi s'inscrit résolument dans une approche de type raisonnement à partir de cas. Pour ce qui me concerne, l'intérêt de ce défi réside dans le fait de traiter des textes sur lesquels se posent des problèmes de représentation d'une recette, de comparaison et de choix d'une recette mais aussi d'adaptation d'une recette pour enlever un ingrédient ou pour substituer un ingrédient par un autre.

Le domaine de la cuisine est riche en ressources textuelles. Mais en premier lieu, il fait appel à une certaine part d'intuitif chez chacun de nous, voire même d'expérience. Mais il s'agit là d'un argument à double tranchant ! S'il nous est plus facile d'interpréter ou d'évaluer un résultat, chacun possède sa propre intuition sur les principes devant guider un tel système, principes qu'il n'est pas toujours facile de partager avec d'autres. De même, l'ontologie culinaire est très dépendante de la culture et, d'une région à l'autre, ou d'un pays à l'autre, des ingrédients rejetés ou peu appréciés par certains seront considérés comme recherchés pour d'autres (on peut penser à l'exemple des œufs de 1 000 ans par exemple).

De nombreuses ressources sont donc disponibles sur le web : bases de recettes, thésaurus, astuces de substitution... Toutes ces ressources peuvent contribuer à une meilleure modélisation du domaine. Sur ce projet, je m'intéresse à deux questions : comment construire une ontologie des ingrédients qui puisse être utilisée pour la substitution d'ingrédients ? Comment définir des sortes de prototypes dans le domaine de la cuisine.

Construire une ontologie des ingrédients

L'idée sous-jacente à une ontologie des ingrédients est de disposer d'une classification des ingrédients qui reflète une certaine "proximité culinaire" entre les ingrédients : un ingrédient pourrait être substitué par un ingrédient de la même classe. De façon similaire à nos travaux dans le domaine de l'astronomie [BTN08], nous pouvons exploiter des ressources catégorisant les ingrédients (`poire est-un fruit`) comme le cook's thesaurus et les nombreuses recettes disponibles sur le web pour caractériser la façon dont un ingrédient peut être préparé.

L'ontologie est utilisée par le moteur de raisonnement à partir de cas (RAPC). La qualité de l'ontologie peut être évaluée au travers des réponses que le moteur de RAPC donne aux requêtes qui lui sont fournies. Taaable 2009 ayant été développé autour d'un wiki sémantique, ce projet soulève des questions ouvertes sur la cohérence d'un système collaboratif et distribué dans lequel machine et humain coopèrent (voir section 5.5.2) ainsi que des questions sur l'évaluation d'une ontologie (voir section 5.5.1).

Construire des prototypes

La notion de prototype apparaît assez naturellement dans un domaine comme la cuisine. L'idée est de définir des objets génériques comme une soupe, une sauce... Qu'est-ce

Rice Soup :
Heat ghee in a large soup pot. Saute diced vegetables and garlic for 15 minutes. Add water, bring to the boil, and simmer until the vegetables are almost tender. While the vegetables are simmering, add the herbs in order listed. Add cooked rice 5 minutes before the end of the cooking time. Add more water, if required.

FIGURE 5.2 – Recette de la soupe de riz

Cream of Vegetable Soup :
Melt butter in a large stock pot over a medium heat. Add onion and saute for 1 to 2 minutes. Reduce heat to low and add remaining ingredients, except stock, cream and parsley. Cook until vegetables are soft, but not brown, for about 20 to 25 minutes. Add stock and bring to the boil over a medium to high heat. Reduce heat and simmer for about 10 minutes. Cook slightly. Transfer to blender or processor in batches and puree. Taste and adjust seasoning. Return to stock pot, place over medium heat and gradually stir in cream. Heat through, but do not boil. Garnish with parsley.

FIGURE 5.3 – Recette de la crème de légume

donc qu'une soupe ? Prenons l'exemple de la soupe de riz (voir Figure 5.2) et de la crème de légume (voir Figure 5.3). Quels sont les points communs entre ces deux soupes et peut-on à partir de ces deux recettes ébaucher le concept de soupe ? Une méthode de comparaison un peu abrupte pourrait faire appel à un analyseur syntaxique produisant des structures prédicat-argument comme le montre la table 5.1, puis d'en faire l'intersection. L'intersection se résume malheureusement à trois prédicats assez peu significatifs. Même si un prétraitement plus soigneux aurait sans doute permis une meilleure analyse des recettes, cette approche très dépendante de la formulation a assez peu de chance d'aboutir.

En revanche, une comparaison phrase par phrase "à la main" (voir figure 5.4) fait ressortir trois étapes communes à ces deux soupes ainsi que des étapes qui sont spécifiques à chacune d'elles. Ces trois étapes communes pourraient probablement être acceptables comme un prototype de soupe.

À un autre niveau, Taaable peut avoir besoin de prototypes. Il est rare que l'on puisse

TABLE 5.1 – Exemple de relations prédicats-arguments extraits de la soupe de riz et, en dernière colonne, liste des prédicats communs aux deux recettes

dobj(Add, water)	det(vegetables, the)	conj_and(Add,bring)
conj_and(Add, bring)	nsubj(tender, vegetables)	det(boil,the)
det(boil, the)	cop(tender, are)	prep_to(bring,boil)
prep_to(bring, boil)	advmod(tender, almost)	
conj_and(Add, simmer)	advcl(simmer, tender)	
mark(tender, until)	...	

Heat ghee in a large soup pot. Melt butter in a large stock pot over a medium heat.
(matière grasse, chaleur, casserole)
Saute diced vegetables and garlic for 15 minutes. Add onion and saute for 1 to 2 minutes. (faire sauter des légumes)
Reduce heat to low and add remaining ingredients, except stock, cream and parsley. Cook until vegetables are soft, but not brown, for about 20 to 25 minutes.
Add water, bring to the boil, and simmer until the vegetables are almost tender. Add stock and bring to the boil over a medium to high heat. Reduce heat and simmer for about 10 minutes. (ajouter liquide, faire bouillir, mijoter)
While the vegetables are simmering, add the herbs in order listed. Add cooked rice 5 minutes before the end of the cooking time. Add more water, if required.
Transfer to blender or processor in batches and puree. Taste and adjust seasoning.
Return to stock pot, place over medium heat and gradually stir in cream. Heat through, but do not boil. Garnish with parsley.

FIGURE 5.4 – Points communs et différences entre les deux recettes

substituer un ingrédient par un autre sans devoir adapter la façon de le préparer. Il est donc intéressant de dégager des recettes, des méthodes prototypiques de préparation des ingrédients et, lors du remplacement d'un ingrédient par un autre, de modifier également le mode de préparation en accord avec les prototypes identifiés. Ce fut l'objet du stage de Valmi Dufour [DLLNT10].

Le domaine de la cuisine peut donc être un domaine d'application très riche pour appliquer des travaux en fouille de textes. La forme linguistique de ces textes – bien que spécifique – est peu complexe relativement à des textes scientifiques. L'extraction d'information permet de se concentrer sur des problèmes qui ne sont pas strictement liés à la forme. Cependant, les deux dimensions du goût et de la texture qui sont essentielles dans la cuisine sont quasi-absents des textes, les résultats seront donc probablement bien en deçà des attentes d'un cuisinier !

5.4.2 Le projet Hybride

Le projet Hybride est une ANR blanche qui débutera en décembre 2011 et dont je suis responsable. Elle implique le LORIA, le GREYC, MoDyCo et Orphanet. Le projet vise à développer de nouvelles méthodes et outils pour la découverte de connaissances à partir de textes, en combinant des méthodes issues du traitement automatique de la Langue (TAL) et de l'extraction de connaissances à partir de bases de données (ECBD). Le domaine d'expérimentation est l'étude des maladies rares. L'idée de ce projet est d'exploiter les méthodes de traitement automatique de la langue pour la fouille de textes et, à l'inverse, d'exploiter les méthodes de fouille de données pour l'analyse de documents textuels. Ainsi, les méthodes du TAL sont utilisées pour identifier des informations dans les textes qui sont utilisées par la suite par les outils d'ECBD. Les méthodes d'ECBD sont utilisées pour extraire des motifs ou des séquences qui sont utilisées pour guider l'extraction d'information.

La documentation d'une maladie rare suppose de lire et de synthétiser un nombre important de textes, puis de compléter les différents champs d'un formulaire définissant les éléments importants pour caractériser une maladie (symptômes, traitements...). Ainsi la phrase suivante, issue d'un article scientifique : "Regardless of patients age and stage of Neuroblastoma disease, amplified expression of MYCN oncogene is the worst paraclinical prognostic factor" établit une relation entre une population de patients, une maladie (neuroblastomie), un gène (oncogène) et un pronostic. Le projet hybride va s'intéresser plus particulièrement, pour un ensemble de maladies, (1) à identifier les gènes qui en sont responsables et le degré de certitude qui peut y être associé et (2) la chronologie des symptômes qui la caractérise pour en faciliter le diagnostic. Le schéma global du projet est donné par la figure 5.5 issue de la proposition du projet.

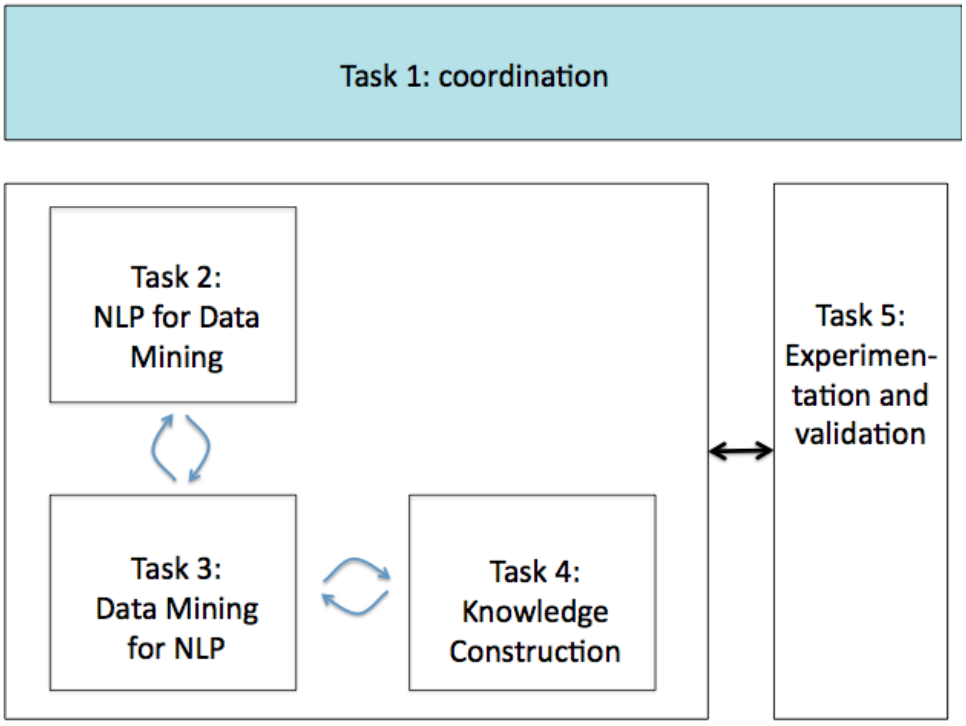


FIGURE 5.5 – Organisation des tâches dans le projet Hybride

Au LORIA, nous nous intéresserons plus particulièrement à modéliser les symptômes sous forme de graphes et à les classifier par l'analyse formelle des concepts. Ainsi, certains graphes de symptômes apparaîtront comme partagés par plusieurs maladies alors que certains autres seront caractéristiques d'une seule maladie.

5.4.3 Le projet KolFlow

Le projet KolFlow s'intéresse à la construction continue de connaissances à partir d'informations en se plaçant dans un espace social et sémantique (un wiki sémantique) où l'homme collabore avec la machine pour produire une connaissance que tous deux puissent utiliser. KolFlow s'intéresse donc à la coévolution du contenu (*i.e.* des informations) et

des connaissances comme étant le résultat de l'interaction homme-machine. Le schéma global du projet est donné par la figure 5.6 issue de la proposition du projet.

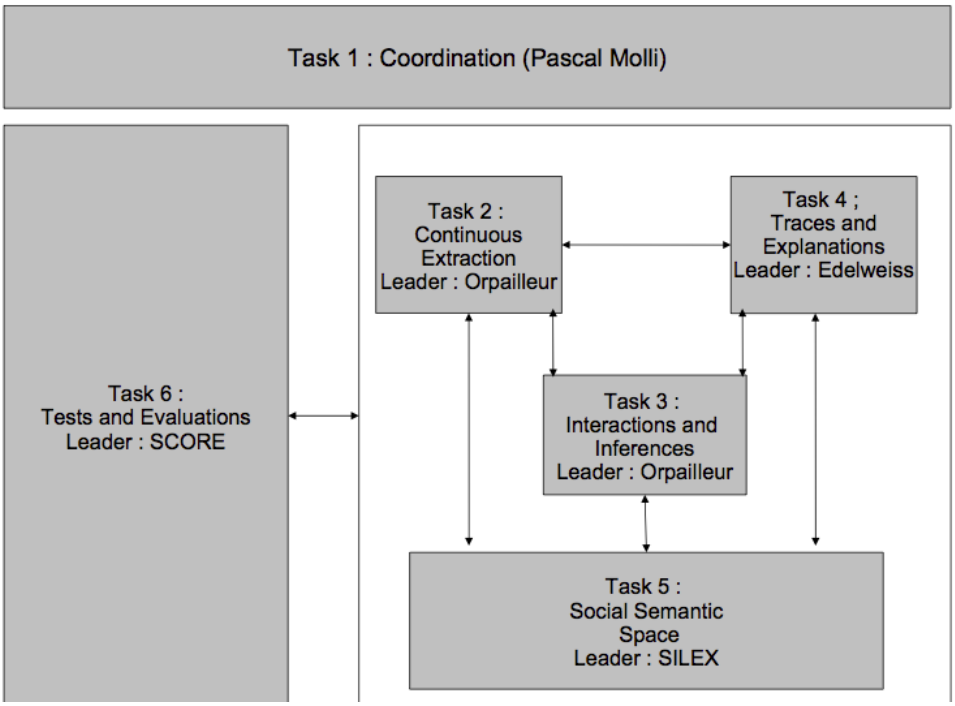


FIGURE 5.6 – Organisation des tâches dans le projet KolFlow

Du point de vue de mon projet de recherche, le point fort de KolFlow vient du fait que le wiki sémantique permet de faire un lien entre les sources textuelles et leur annotation sémantique, l'ontologie – construite par la FCA –, et l'utilisation par un mécanisme de raisonnement à partir de cas de l'ontologie, processus qui peut, en retour, impliquer des modifications dans l'ontologie.

5.5 Eléments de réflexion

Deux questions incontournables pour lesquels il existe encore très peu de travaux fondamentaux doivent être envisagées à court terme. La première concerne l'évaluation de la qualité d'une ontologie qui doit porter à la fois sur une démarche méthodologique solide, voire formelle, mais également prendre en compte des aspects plus subjectifs liés à la tâche ou à l'expert. La seconde est qu'il faut profiter de travaux actuels sur le coopératif pour se poser la question de faire coopérer les experts avec la machine tout en préservant la qualité formelle, par exemple, des ontologies produites. Enfin, le dernier point de cette section porte sur la diffusion de nos travaux en dehors du laboratoire par la réalisation d'une plateforme de fouille de textes.

5.5.1 Evaluation d'une ontologie

Que l'on considère les ontologies comme la capitalisation des connaissances sur un domaine ou comme une synthèse partielle, voire locale, d'un sous-domaine, la question de l'évaluation du résultat, ou plus généralement de qualité de l'ontologie produite, est un problème complexe. Il n'existe pas de "gold standard" utilisable pour confronter nos résultats à cette référence. La question que je me pose ici est donc de se donner les moyens d'apprécier la qualité de l'ontologie produite, tant par rapport aux textes que par rapport aux objectifs d'un expert qui piloterait cette construction. Je ferai référence ici à deux types de travaux. Les premiers travaux [GP04] portent sur les propriétés que doit vérifier une ontologie. Les seconds, [GW00b, GW02, GW04], portent sur la nature des informations permettant d'organiser, de structurer et de raisonner sur une ontologie.

[GP04] définit les propriétés attendues d'une ontologie telles **la cohérence, la complétude, la concision, l'expansivité et la sensibilité**. La construction de treillis peut être réalisée par un processus incrémental, la complétude pourrait être reflétée par la façon dont le treillis évolue relativement à la quantité de textes traités. La concision pourrait être vue au travers d'indices statistiques comme la stabilité, bien qu'il soit probable que ce seul indice ne puisse refléter l'intérêt d'un concept du treillis pour l'expert.

La FCA et la RCA sont des méthodes qui génèrent un très grand nombre de classes. Toutes ces classes ne sont pas pertinentes en tant que classe pour une ontologie. [GW00b] définit les méta-propriétés de rigidité, d'unité, d'identité et de dépendance. L'objectif de ces travaux est de montrer que ces propriétés peuvent conduire à des types de subsomption différents. Je reprends ici quelques éléments de définition qui sont proposées par les auteurs :

- la **rigidité** est une propriété essentielle pour toutes les instances. A l'inverse la non-rigidité pour une propriété est le fait d'être non essentielle pour certaines instances d'une classe. La propriété anti-rigide est une propriété qui n'est pas essentielle pour l'ensemble des instances et la propriété semi-rigide est une propriété non-rigide mais pas anti-rigide. La notion de rigidité et d'identité peut-être lié à la diachronie. Ainsi, un étudiant est une personne mais un individu peut cesser d'être étudiant mais ne cessera pas d'être une personne.
- l'**unité** est définie [GW00b] comme la capacité à distinguer les parties d'une instance du reste du monde par la relation d'unification qui les relie ensemble. Une question relative à l'unité serait "est-ce que le collier est une partie de mon chien ?"
- l'**identité** concerne la différenciation, au sein d'une même classe de l'ontologie, des différents individus. Cette différenciation peut se faire par le biais d'attributs ou de relations qui lui sont uniques. Une question du type "Est-ce mon chien ?" est une question d'identité.
- la **dépendance** s'explique par la nécessité qu'une autre entité existe. Par exemple, `Parent` dépend de l'existence de `Enfant`.

Ces définitions qui restent très théoriques dans [GW00b] ont été mises en œuvre dans [VVSH08] par la définition de patrons lexico-syntaxiques.

L'idée que je souhaite développer dans ce travail est que les informations que l'on extrait des textes n'ont pas la même valeur "ontologique". Les concepts, les attributs et les relations sont actuellement identifiés *a priori* en définissant le modèle de l'ontologie. L'uti-

lisation des méta-propriétés pourrait y contribuer de façon moins arbitraire et permettrait d'associer à une information un statut qui pourrait être introduit dans le processus de fouille.

5.5.2 Interaction avec l'expert et wikis sémantiques

La meilleure évaluation de l'ontologie reste probablement d'évaluer si l'ontologie répond à la tâche pour laquelle elle a été construite et si elle peut être enrichie ou corrigée au cours du temps. La section 3.4.4 présente dans le contexte de la FCA et de la RCA un ensemble d'opérations de base permettant de revenir sur les données d'entrée, de les corriger ou de les enrichir pour qu'au final l'ontologie soit mieux en accord avec les attentes de l'expert.

Taaable 2009 a été développé autour d'un wiki sémantique (Semantic MediaWiki). Les wikis sémantiques proposent un environnement collaboratif, voire distribué, il semble donc naturel d'exploiter cet environnement pour représenter l'ontologie et permettre aux experts d'interagir et d'en modifier le contenu. Ces wikis sont dotés d'une syntaxe et d'une sémantique proche de celles d'une logique de descriptions : [KSV07] montre qu'il est possible de mettre en correspondance les annotations des wikis sémantiques avec les logiques de descriptions.

Dans [Ben09], l'idée que nous défendions était que la phase critique de la construction d'une ontologie est le passage à une structure formelle. De ce point de vue, la FCA et la RCA apportent une solution très performante. En revanche, cette idée a aussi conduit à ce que l'interaction de l'expert se fasse au niveau des données et un nouvel appel à la FCA-RCA recalcule la nouvelle ontologie. L'intérêt et les inconvénients de cette approche peuvent se résumer par quelques points :

+ la cohérence de l'ontologie est assurée de par le principe de construction avec la FCA ;
+ l'historique des modifications peut être sauvegardé et le processus est donc réappliqué aux mêmes données et dans les mêmes conditions ;
- les opérations proposées sur les données sont très élémentaires et nécessitent souvent de passer en revue la description de tous les objets du domaine ;
- l'expert peut être insatisfait du treillis. Il peut avoir envie de voir émerger une classe ou de voir disparaître une certaine autre classe sans pour autant savoir comment corriger les données.

L'interaction avec l'expert doit passer par la définition d'opérations mieux appropriées que les opérations actuellement accessibles. Des environnements comme Protégé [7] ont été développés pour faciliter l'édition d'ontologie et leur association à un raisonneur comme Pellet [PS04] permettent d'en vérifier la cohérence. Dans un environnement collaboratif et distribué, les wikis sémantiques sont des environnements dédiés à la prise en compte de l'utilisateur, i.e. de l'expert.

Le point que je souhaite aborder dans un tel contexte porte sur comment articuler la création automatique d'ontologie à partir de ressources textuelles et interaction humaine. Dans un tel contexte, les points sur lesquels je souhaite me concentrer portent

7. Protégé : http ://protege.stanford.edu/, dernière visite le 05/10/2009

sur la cohérence des ontologies produites et la possibilité d'enrichir de telles ontologies. Un certain nombre de travaux existent actuellement sur la fusion d'ontologies. J'ai cité précédemment [SM01b] qui propose une fusion guidée par les observations en corpus mais il faut également se référer aux travaux menés dans le cadre du projet européen NEON.

5.5.3 Construction d'une plateforme de Text Mining

L'objectif de cette plateforme « Knowledge Discovery from Texts (KnoTs) » est de développer un environnement de fouille de textes générique. Cette plate-forme complète de fouille de textes, robuste et paramétrable, doit nous permettre plusieurs choses : de tester nos nouvelles idées en matière de fouille de textes et d'ouvrir de nouvelles directions de recherche, de conforter les collaborations de l'équipe sur les plans national et international (participation à des projets de recherche pour les appels d'offres ANR et européens) mais aussi de gagner en visibilité. Enfin, cette plate-forme doit être suffisamment ouverte pour être facilement enrichie des derniers travaux de l'équipe. L'équipe possède un savoir-faire sur le plan théorique mais qui n'a été validé jusqu'à présent que par des expériences locales et de faible volume. La production de connaissances à partir de contenus est en plein essor et une plate-forme "prête à l'emploi" traitant l'activité de fouille de textes nous donnerait un avantage et une visibilité indéniables. Il faut penser les choses à l'échelle du Web et, autrement dit, il faut pouvoir passer à l'échelle et valider sur de gros corpus de données ce que nous savons faire sur des faibles volumes de données. Le système développé sera considéré comme un module qui pourra s'intégrer à la plate-forme GATE : l'architecture logicielle robuste de cet environnement, sa très grande modularité et sa popularité dans le monde du traitement de documents seront des atouts pour le développement et la visibilité de nos travaux. Un ingénieur a été affecté à cette tâche en novembre 2010 pour intégrer nos travaux.

Chapitre 6

Conclusion

Il n'existe pas d'outils clé en main pour extraire des connaissances de textes et nous avons montré que le passage de la langue naturelle à des connaissances est très fortement contextualisé et dépendant de la tâche que l'on s'est fixée. Nous avons vu que le défi est vaste et les pistes de recherche très nombreuses, que ce soit dans la recherche d'information, dans le traitement automatique des langues, dans la fouille de données ou dans la représentation des connaissances. Chacun de ces domaines de recherche recensent de nombreux sous-domaines tous très actifs.

Le projet de recherche que je souhaite développer peut être vu comme un chemin au travers de ces domaines qui vise à créer un continuum (sémantique) entre les différentes étapes de la fouille de textes. L'extraction de connaissances à partir de textes est avant tout une construction de connaissances et suppose une cohérence méthodologique entre les différentes étapes de la fouille de textes.

J'ai fait le choix d'ancrer mes travaux dans le domaine du formel en visant notamment une représentation du sens en logique, plus particulièrement en logique de descriptions. Malgré les restrictions liées à ce choix, notamment en ce qui concerne l'interaction avec des humains experts d'un domaine, la mise à jour, ou la correction d'une ontologie, une représentation formelle reste à mon sens la solution pour raisonner sur les textes et assurer la cohérence d'une ontologie.

Si le but final d'un processus de fouille est de construire une représentation formelle qui puisse être le support de raisonnements, je me suis concentré dans ce projet de recherche sur la construction des connaissances en exploitant des méthodes à base de motifs, d'extraction de règles d'association ou de l'analyse formelle de concepts. L'intérêt de ces approches est qu'elles assureront un lien constant entre les textes et les connaissances. La modification des textes engendre une modification des connaissances et inversement la modification des connaissances (les ressources externes par exemple) modifient l'annotation des textes et l'ontologie. Des environnements coopératifs comme les wikis sémantiques pourraient à terme intégrer nos travaux et faciliter ainsi la synergie entre les processus humains et les processus automatiques.

Bibliographie

[AG05] N. Aussenac-Gilles. *Méthodes ascendantes pour l'ingénierie des connaissances* . Habilitation à diriger des recherches, Université Paul Sabatier, Toulouse, France, décembre 2005.

[AIS93] R. Agrawal, T. Iielinski, and A. Swami. Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD Conférence on Management of Data*, pages 207–216, Washington, D.C., 1993.

[Ama07] F. Amardeilh. *Web sémantique et informatique linguistique : propositions méthodologiques et réalisation d'une plateforme logicielle*. PhD thesis, Thèse de doctorat, Univ. Paris X, 2007.

[AMS+96] R. Agrawal, H. Mannila, R. Shrikant, H. Toivonen, and A.I. Verkamo. Fast discovery of association rules. In Fayyad et al. [FPSSU96], pages 307–328. 625 pages.

[AR03] J. Aze and M. Roche. Une application de la fouille de textes : l'extraction des règles d'association à partir d'un corpus spécialisé. *Revue d'intelligence artificielle - Acte des Journées sur l'Extraction et la Gestion des Connaissances*, 17(1-3) :283–294, 2003.

[Arm94] S. Armstrong, editor. *Using Large Corpora*. MIT Press, 1994.

[AS94] R. Agrawal and R. Shrikant. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on verry large Databases (VLDB'94)*, pages 487–499, San Francisco, CA, USA, 1994. Morgan Kauffmann Publisher Inc.

[Azé03] J. Azé. *Extraction de connaissances à partir de données numériques et textuelles*. PhD thesis, Université Paris-Sud 11 et laboratoire de Recherche en Informatique, UMR C.N.R.S. 8623, Orsay, 2003.

[Bar66] R. Barthes. *Critique et vérité*. Le Seuil, Paris, 1966.

[BBB+08] F. Badra, R. Bendaoud, R. Bentebibel, P.-A. Champin, J. Cojan, A. Cordier, S. Després, S. Jean Daubias, J. Lieber, T. Meilender, A. Mille, E. Nauer, A. Napoli, and Y. Toussaint. TAAABLE : Text Mining, Ontology Engineering, and Hierarchical Classification for Textual Case-Based Cooking. In Martin Schaaf, editor, *9th European Conference on Case-Based Reasoning - ECCBR 2008*, pages 219–228, Trier Allemagne, 2008.

[BBC+92] M. Borillo, A. Borillo, N. Castell, D. Latour, and Y. Toussaint. Applying linguistic engineering to software engineering : The traceability problem. In *Actes de European Conference on Artificial intelligence (ECAI)*, Vienne, Autriche, Août 1992.

[BCC+09] F. Badra, J. Cojan, A. Cordier, J. Lieber, T. Meilender, A. Mille, P. Molli, E. Nauer, A. Napoli, H. Skaf-Molli, and Y. Toussaint. Knowledge acquisition and discovery for the textual case-based cooking system taaable. In *actes de 2nd Computer Cooking Contest en association avec International Conference on Case Based Reasoning*, Seattle, USA, juillet 2009. To be published.

[BCD+90] M. Bras, D. Coulon, J.-P. Desclès, C. Fuchs, F. Gayral, J. Jayez, D. Kayser, F. Nef, D. Reppert, P. Saint-Dizier, C. Tollu, Y. Toussaint, and B. Victorri. La sémantique des langues naturelles : éléments d'une approche comparative. In Hermes, editor, *actes des 3èmes journées du PRC-GDR Intelligence Artificielle*, 1990.

[BCM+02] F. Baader, D. Calvanese, D. McGuinness, D. Nardi, and P. F. Patel-Schneider, editors. *The Description Logic Handbook : Theory, Implementation and Applications*, chapter Basic Description Logics. Cambridge University Press, 2002.

[BCM05] P. Buitelaar, P. Cimiano, and B. Magnini, editors. *Ontology learning from text : methods, evaluation and applications*. Frontiers in Artificial Intelligence and Applications. IOS Press, 2005.

[Ben66] E. Benveniste. Formes nouvelles de la composition nominale. *Bulletin de la Société de Linguistique de Paris*, 61 :82–95, 1966.

[Ben04] R. Bendaoud. Fouille de données textuelles complexes. Master's thesis, DEA de l'Université Henry Poincaré – nancy I, spécialité Informatique, 2004.

[Ben09] R. Bendaoud. *Analyse formelle et Relationelle de concepts pour la construction d'ontologies de domaines à partir de ressources textuelles héterogènes*. PhD thesis, LORIA – Université Henri Poincarré Nancy 1, juillet 2009.

[Ber04] M. Berry, editor. *Survey of text mining : Clustering, classification and retrieval*. Springer Verlag, 2004.

[Bib92] D. Biber. The multidimensional approach to linguistic analyses of genre variation : An computers in the humanities. *Computers and the humanities*, 26(5-6) :331–347, 1992.

[Bir97] G. Birkhoff. *Lattice Theory*. American Mathematical Society Colloquium Publications, New York, 3rd ed. edition, 1997.

[BLC96] A. Bernaras, I. Laresgoiti, and J. Corera. Building and reusing ontologies for electrical network applications. In *12th European Conference on Artificial Intelligence*, pages 298–302, Budapest, 1996.

[BM70] M. Barbut and B. Monjardet. *Ordre et classification*, volume 2 tomes. Hachette, Paris, 1970.

[BMSW97] D. M. Bikel, S. Miller, R. Schwartz, and R. Weischedel. Nymble : a high-performance learning name-finder. In *Proceedings of the fifth conference on Applied natural language processing*, pages 194–201, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc.

[BMUT97] S. Brin, R. Motwani, J. D. Ullman, and S. Tsur. Dynamic itemset counting and implication rules for market basket data. In *Proceedings of the ACM SIGMOD international conference on Management of data*, volume 26 of *ACM SIGMOD*, 1997.

[BNT08a] R. Bendaoud, A. Napoli, and Y. Toussaint. A Proposal for an Interactive Ontology Design Process based on Formal Concept Analysis. In Carole Eschenbach and Michael Grüninger, editors, *5th international conference on Formal Ontology in Information Systems - FOIS 2008*, volume 183, pages 311–323, Saarbrücken Allemagne, 2008. IOS Press.

[BNT08b] R. Bendaoud, A. Napoli, and Y. Toussaint. Formal Concept Analysis : A unified framework for building and refining ontologies. In Aldo Gangemi and Jérôme Euzenat, editors, *16th International Conference on Knowledge Engineering and Knowledge Management - EKAW 2008*, volume 5268, pages 156–171, Acitrezza, Catania Italie, 2008. Springer Berlin / Heidelberg.

[Bor86] J-P. Bordat. Calcul pratique du treillis de galois d'une correspondance. *Mathématiques et Sciences Sociales*, 96 :37–41, 1986.

[Bor09] C. Borgelt. Graph Mining : An Overview. In *Proceedings of the 19th GMA/GI Workshop Computational Intelligence*, pages 189–203, 2009.

[Bou94] D. Bourigault. *LEXTER, un Logiciel d'EXtraction de TERminologie. Application à l'acquistion des connaissances à partir de textes.* PhD thesis, Ecole des Hautes Etudes en Sciences Sociales, June 1994.

[BRHT+07a] R. Bendaoud, M. Rouane Hacene, Y. Toussaint, B. Delecroix, and A. Napoli. Construction d'une ontologie à partir d'un corpus de textes avec l'ACF. In *Ingénierie des Connaisances IC 2007*, Grenoble France, juillet 2007.

[BRHT+07b] R. Bendaoud, M. Rouane Hacene, Y. Toussaint, B. Delecroix, and A. Napoli. Text-based ontology construction using relational concept analysis. In *International Workshop on Ontology Dynamics - IWOD 2007*, Innsbruck Autriche, 06 2007.

[Bri92] E. Brill. A simple rule-based part of speech tagger. In *Proceedings of the third conference on Applied natural language processing, ANLP'92*, pages 152–155, 1992.

[Bri93] E. Brill. *A Corpus-Based Approach to Language Learning.* PhD thesis, University of Pennsylvania, 1993.

[BS99] B. Biébow and S. Szulman. Terminae : A linguistic-based tool for the building of a domain ontology. In *Proceedings of the EKAW Conference (EKAW'99)*, number 1621 in Lecture Notes in Artificial Intelligence (LNAI), pages 49–66. Springer, 1999.

[BT93] M. Bras and Y. Toussaint. Artificial intelligence tolls for software engineering. In *8th international Conference on Applications of Artificial Intelligence in Engineering (AIENG'1993)*, Toulouse, 1993.

[BT00] A. Belaïd and Y. Toussaint. Une méthode d'étiquetage morpho-syntaxique pour la reconnaissance de tables de matières. In *Colloque International Francophone sur l'Ecrit et le Document - CIFED'00*, page 10 p, Lyon, France, juillet 2000.

[BTB91a] M. Borillo, Y. Toussaint, and A. Borillo. Motivations du projet lesd. In *actes de la conférence "Linguistique Engineering'91"*, Janvier 1991.

[BTB91b] M. Borillo, Y. Toussaint, and A. Borillo. Motivations du projet linguistic engineering for software design (lesd). *Génie Logiciel et Systèmes Experts, numéro spécial "Génie Logiciel et Langage naturel"*, 23 :78–83, juin 1991.

[BTBS92] M. Bras, Y. Toussaint, M. Borillo, and A. Simon. Towards‘ an electronic dictionary of the space technical domain. In *World Space Congress*, Washington, 1992. IAF and COSPAR.

[BTN05] R. Bendaoud, Y. Toussaint, and A. Napoli. Hiérarchisation des règles d'association en fouille de textes. *Revue des Sciences et Technologies de l'Information (Série Ingénierie des Systèmes d'Information)*, 1 :263–274, 2005.

[BTN07] R. Bendaoud, Y. Toussaint, and A. Napoli. Construction d'ontologie à partir de corpus de textes. In *Septièmes journées Extraction et Gestion des Connaissances - EGC 2007*, Namur Belgique, 01 2007.

[BTN08] R. Bendaoud, Y. Toussaint, and A. Napoli. PACTOLE : A methodology and a system for semi-automatically enriching an ontology from a collection of texts. In *16th International Conference on Conceptual Structures ICCS'08*, volume 5113, pages 203–216, Toulouse France, 2008.

[BTN10] R. Bendaoud, Y. Toussaint, and A. Napoli. L'analyse Formelle de Concepts au service de la construction et l'enrichissement d'une ontologie. *Revue des Nouvelles Technologies de l'Information*, E-18 :133–164, February 2010.

[BTP+02] Y. Bastide, R. Taouil, N. Pasquier, G. Stumme, and L. Lakhal. Un algorithme d'extraction des motifs fréquents. *Techniques et Sciences informatiques*, 21(1) :65–95, 2002.

[CF06] D. Chakrabarti and C. Faloutsos. Graph mining : Laws, generators, and algorithms. *ACM Computing Surveys*, 38, 2006.

[CFH06] S. Castano, A. Ferrara, and G. N. Hess. Discovery-driven ontology evolution. In G. Tummarello, P. Bouquet, and O. Signore, editors, *3rd Italian Semantic Web Workshop*, Pisa, Italy, 2006.

[CH07] D.J. Cook and L.B. Holder, editors. *Mining Graph Data*. John Wiley & Sons, Hoboken (NJ), 2007.

[Che04] H. Cherfi. *Etude et réalisation d'un système d'extraction de connaissances à partir de textes*. PhD thesis, Université Henri Poincaré - Nancy 1, spécialité informatique, 2004.

[CHS05] P. Cimiano, A. Hotho, and S. Staab. Learning concept hierarchies from text corpora using formal concept analysis. In *Journal of Artificial Intelligence Research (JAIR'05)*, volume Volume 24, pages 305–339. AAAI Press, 2005.

[Cim05] P. Cimiano. *Ontology Learning and Population from Text : Algorithms, Evaluation and Applications*. Springer, 2005.

[CJNT04] H. Cherfi, D. Janetzko, A. Napoli, and Y. Toussaint. Sélection de règles d'association par un modèle de connaissances pour la fouille de textes. In M. Liquière et M. Sebban, editor, *Conférence d'Apprentissage - CAp 2004*, pages 191–206, Montpellier, France, 2004. Presses Universitaires de Grenoble.

[CK99] M. Craven and J. Kumlien. Constructing biological knowledge bases by extracting information from text sources. In *Proceeding of the 7th International Conference on Intelligent Systems for Molecular Biology (ISMB-99)*, pages 77–86, Heidelberg, Germany, 1999. AAAI Press.

[CL91] K. W. Church and M. Y. Liberman. A status report on the acl/dci. In *Proceedings of the 7th Annual Conference of the UW Centre for the New OED and Text Research : Using Corpora*, pages 84–91, Oxford, 1991.

[CLM+09] A. Cordier, J. Lieber, P. Molli, E. Nauer, H. Skaf-Molli, and Y. Toussaint. WikiTaaable : A semantic wiki as a blackboard for a textual case-based reasoning system. In *4th Workshop on Semantic Wikis (SemWiki2009), held in the 6th European Semantic Web Conference*, May 2009.

[CLNT09] A. Cordier, J. Lieber, E. Nauer, and Y. Toussaint. Taaable, une application culinaire du raisonnement à partir de cas. In *Session industrie et démonstrations de EGC'09 (Extraction et Gestion des Connaissances).*, January 2009.

[CM98] L. Chaudron and N. Maille. First order logic formal concept analysis : from logic programming to theory. *Computer and Information Science*, 13(3), 1998.

[CNT00a] F. Chakkour, A. Napoli, and Y. Toussaint. Le raisonnement à partir de cas pour l'identification de rôles sémantiques dans des énoncés en langue naturelle. In *séminaire RàPC-2000*, 05 2000.

[CNT00b] N. Collier, C. Nobata, and J.-I. Tsujii. Extracting the names of genes and gene products with a hidden markov model. In *Proceedings of the 18th conference on Computational linguistics*, pages 201–207, Morristown, NJ, USA, 2000. Association for Computational Linguistics.

[CNT01] F. Chakkour, A. Napoli, and Y. Toussaint. Extraire des structures prédicatives à partir des textes, vers une indexation conceptuelle des textes. In Alain Mille Béatrice Fuchs, editor, *Plate-forme AFIA / Atelier raisonnement partir de cas*, page 7 p, Grenoble, France, 2001.

[CNT03a] H. Cherfi, A. Napoli, and Y. Toussaint. Towards a Text Mining Methodology Using Frequent Itemsets and Association Rule Extraction. In M.

Nadif, A. Napoli, E. SanJuan, and A. Sigayret, editors, *Journées d'informatique Messine - JIM'03*, pages 285–294, Metz, France, septembre 2003. INRIA Lorraine.

[CNT03b] H. Cherfi, A. Napoli, and Y. Toussaint. Vers une méthodologie de fouille de textes s'appuyant sur l'extraction de motifs fréquents et de règles d'association. In Rémi Gilleron, editor, *Conférence d'Apprentissage (CAp'03), dans le cadre de la plate-forme (AFIA'03)*, pages 61–76, Laval, France, juillet 2003. Presses universitaires de Grenoble.

[CNT05a] H. Cherfi, A. Napoli, and Y. Toussaint. Deux méthodes de classification de règles d'association en fouille de textes. In V. Makarenkov ; G. Cucumel ; F.-J. Lapointe, editor, *12èmes journées de la Société Francophone de Classification - SFC-05*, pages 104–107, Montréal/Canada, 04 2005. Presses Universitaires de Montréal.

[CNT05b] H. Cherfi, A. Napoli, and Y. Toussaint. Deux méthodologies de classification de règles d'association pour la fouille de textes. *Revue Des Nouvelles Technologies de l'Information*, E-4 :211–234, 2005.

[CNT06] H. Cherfi, A. Napoli, and Y. Toussaint. Towards a Text Mining Methodology Using Association Rules Extraction. *Soft Computing*, 10 :431–441, 2006.

[CNT09] H. Cherfi, A. Napoli, and Y. Toussaint. *Post-Mining of Association Rules : Techniques for Effective Knowledge Extraction*, chapter A Conformity Measure using Background Knowledge for Association Rules : Application to Text Mining (Chap. 6), pages 100 – 115. IGI Global, yanchang zhao and chengqi zhang and longbing cao edition, mai 2009.

[Con03] A. Condamines. Sémantique et corpus spécialisés : Constitution de bases de connaissances terminologiques. Thèse d'habilitation à diriger des recherches, Equipe de Recherche en Syntaxe et Sémantique, Université de Toulouse II - Le Mirail, 2003.

[CR96] C. Carpineto and G. Romano. Information retrieval through hybrid navigation of lattice representations. *International Journal of Computer Human Studies*, 45(5) :553–578, 1996.

[CR98] C. Carpineto and G. Romano. Effective reformulation of boolean queries with concept lattices. In *Flexible Query Answering Systems, Third International Conference (FQAS'98)*, number 1495 in LNCS, pages 83–94. Springer, 1998.

[CR00] C. Carpineto and G. Romano. Order-theoretical ranking. *Journal of the American Society for Information Science*, 51(7) :587–601, 2000.

[CR04] C. Carpineto and G. Romano. Exploiting the potential of concept lattices for information retrieval with credo. *Journal of Universal Computer Science*, 10(8) :985–1013, 2004.

[CS04] A. Culotta and J. Sorensen. Dependency tree kernels for relation extraction. In *Proceedings of the 42nd Meeting of the Association for Computational*

Linguistics (ACL'04), Main Volume, pages 423–429, Barcelona, Spain, July 2004.

[CSHT04] P. Cimiano, G. Stumme, A Hotho, and J. Tane. Conceptual knowledge processing with formal concept analysis and ontologies. In *Proceedings of the The Second International Conference on Formal Concept Analysis (ICFCA 04)*, volume 2961 of *LNAI*, pages 189–207, Sydney, Australia, 2004.

[CSTT94a] N. Castell, O. Slavkova, Y. Toussaint, and A. Tuells. Control de qualidad de las aspecificaciones de software excritas en lenguaje natural. *Novatica, Inteligencia artificial*, 2(109) :33–40, mai 1994.

[CSTT94b] N. Castell, O. Slavkova, A. Tuells, and Y. Toussaint. Quality control of software specification written in natural language. In *Actes de International Conference on Industrial, Engineering & Other Applications of Applied Intelligent SystemsIEA-AIE*, 1994.

[CT96] N. Capponi and Y. Toussaint. Construction et structuration d'un lexique sémantique. In *actes de la journée du Programme de recherches Coordonnées Communication Homme-Machine (PRC CHM)*, 1996.

[CT00] N. Capponi and Y. Toussaint. Interprétation de classes de termes par généralisation de structures prédicat-arguments. In G. Kassel et D. Bourigault J. Charlet, M.Zacklad, editor, *Ingénierie des connaissances, évolutions récentes et nouveaux défis*, pages 337–357. Eyrolles, 2000.

[CT01a] F. Chakkour and Y. Toussaint. Sentence Analysis by Case-Based Reasoning. In Moonis Ali, editor, *The Fourteenth International Conférence on Industrial and Engineering Applications of Artificial Intelligence and Expert SystemsIEA/AIE*, volume 2070, pages 546–551, Budapest, Hungary, 2001. Springer Verlag.

[CT01b] H. Cherfi and Y. Toussaint. Extraction et interprétation de règles d'association pour la fouille de textes. In *A3CTE - Plate-forme AFIA*, pages 15–16, Grenoble, France, 2001. Presses Universitaires de Grenoble.

[CT02a] H. Cherfi and Y. Toussaint. Adéquation d'indices statistiques à l'interprétation de règles d'association. In P. Sébillot A. Morin, editor, *6èmes Journées internationales d'Analyse statistique des Données Textuelles - JADT 2002*, pages 233–244, Saint-Malo, France, 2002.

[CT02b] H. Cherfi and Y. Toussaint. Fouille de textes par combinaison de règles d'association et d'indices statistiques. In *1er Colloque International sur la Fouille de Textes - CIFT'2002*, pages 67–80, Hammamet, Tunisie, 09 2002.

[CT02c] H. Cherfi and Y. Toussaint. How Far Association Rules and Statistical Indices help Structure Terminology ? In A. Maedche N. Aussenac Gilles, editor, *Workshop of ECAI2002 : Natural Language Processing and Machine Learning for Ontology Engineering OLT'02*, pages 5–9, Lyon, France, 07 2002.

[CT02d] H. Cherfi and Y. Toussaint. How far association rules and statistical indices help structure terminology ? In *Workshop on Machine Learning and Natural*

Language Processing for Ontology Engineering held in conjunction with the ECAI'02 conference, Lyon, France, 2002.

[CT02e] H. Cherfi and Y. Toussaint. Interprétation des règles d'association extraites par un processus de fouille de textes. In *13ème Congrès francophone AFRIF-AFIA de Reconnaissance des Formes et d'intelligence Artificielle - RFIA'02*, volume 3, pages 975–983, Angers, France, 2002.

[CT03] H. Cherfi and Y. Toussaint. Méthodologie de sélection et de lecture de règles d'association pour la fouille de textes. In *Atelier de fouille de textes en Génomique, dans le cadre de la conférence Extraction et de Gestion des Connaissances - EGC'03*, pages 1–2, Lyon, France, 01 2003.

[CV05] P Cimiano and J Völker. Text2onto - a framework for ontology learning and data-driven change discovery. In Springer, editor, *Proceedings of Natural Language Processing and Information Systems (NLDB 2005)*, number 3513 in Lecture Notes in Computer Science (LNCS), pages 227–238, June 2005.

[Dai94] B. Daille. *Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques*. PhD thesis, Université de Paris VII, Computer Communication and Vision, TALANA, février 1994.

[Dai02] B. Daille. *Découvertes linguistiques en corpus*. PhD thesis, Mémoire d'Habilitation à Diriger des Recherches en Informatique, Université de Nantes, 2002.

[DLLNT10] V. Dufour-Lussier, J. Lieber, E. Nauer, and Y. Toussaint. Text adaptation using formal concept analysis. In *Proceedings of the International Conference on Cas based Reason ing (ICCBR)*, 2010.

[DMBSV02] M. Delgado, M. Martin-Bautista, D. Sanchez, and M.-A. Vila. Mining text data : Special features and patterns. In D. Hand, N. Adams, and R. Bolton, editors, *Pattern Detection and Discovery, Proceedings of ESF Exploratory Workshop*, volume 2447 of *Lecture Notes in Artificial Intelligence - LNAI*, pages 140–153, 2002.

[DP90] S. David and P. Plante. De la nécessité d'une approche morpho-syntaxique dans l'analyse des textes. *Intelligence artificielle et sciences cognitives au Québec*, 3(3) :140–154, 1990.

[DP07] J. Ducrou and P.Eklund. Searchsleuth : The conceptual neighbourhood of an web query. In J. Diatta, P. Eklund, and M. Liquiere, editors, *Fifth International Conference on Concept Lattices and Their Applications (CLA2007)*, pages 253–263, 2007.

[DRP00] B. Daille, J. Royauté, and X. Polanco. Evaluation d'une plate-forme d'indexation de termes complexes. *Traitement Automatique des Langues (TAL)*, 41(2) :395–422, 2000.

[DSDST89] Christophe D., P. Saint-Dizier, P. Sébillot, and Y. Toussaint. *Logic Grammars and Logic Programming*, chapter A Language Processing System Based on Principles of Government and Binding Theory. Ellis Harwood, 1989.

[EMTB08] A. Estacio Moreno, Y. Toussaint, and C. Bousquet. Mining for adverse drug events with formal concept analysis. *Studies in health technology and informatics*, 136 :803–8, 2008.

[EP95] C. Enguehard and L. Pantera. Automatic natural acquisition of terminology. *Journal of Quantitative Linguistics*, 2(1) :27–32, 1995.

[FD95a] R. Feldman and I. Dagan. Knowledge discovery in texts. In *Proceedings of the ECML-95 Workshop on Knowledge Discovery*, pages 175–180, 1995.

[FD95b] R. Feldman and I. Dagan. Knowledge discovery in textual databases (kdt). In *Proceedings of the 1st International Conference on Knowledge Discovery (KDD-95)*, pages 112–117, Montreal, 1995.

[Fen07] D. R. Feno. *Mesures de qualité des règles d'association : normalisation et caractérisation des bases*. PhD thesis, Université de la Réunion, 2007.

[Fer02] S. Ferré. *Systèmes d'information logiques : un paradigme logico-contextuel pour interroger, naviguer et apprendre*. PhD thesis, Université de Rennes 1, oct 2002. Accessible en ligne depuis l'adresse http ://www.irisa.fr/LIS/ferre/.

[FF10] A. Foret and S. Ferré. On categorial grammars as logical information systems. In L. Kwuida and B. Sertkaya, editors, *Int. Conf. Formal Concept Analysis*, LNCS 5986, pages 225–240. Springer, 2010.

[FFK+98] R. Feldman, M. Fresko, Y. Kinar, Y. Lindell, O. Liphstat, M. Rajman, Y. Schler, and O. Zamir. Text mining at the term level. In *Proceedings of the Second European Symposium on Principles of Data Mining and Knowledge Discovery*, volume 1510 of *Lecture Notes In Computer Science*, pages 65–73. Springer-Verlag London, UK, 1998.

[FH96] R. Feldman and H. Hirsh. Mining associations in text in the presence of background knowledge. In *Knowledge Discovery and Data Mining*, pages 343–346, 1996.

[FH97] R. Feldman and H. Hirsh. Exploiting background information in knowledge discovery from text. *Journal of Intelligent Information Systems*, 9 :83–97, 1997.

[FNR98] D. Faure, C. Nedellec, and C. Rouveirol. Acquisition of semantic knowledge using machine learning methods : the system asium. Technical report, ICS-TR-88-16, LRI, Université Paris Sud, 1998.

[For] R. Forsyth. Uci machine learning repository. http ://archive.ics.uci.edu/ml/datasets.html, dernière visite le 14 septembre 2009.

[FPA06] G. Flouris, D. Plexousakis, and G. Antoniou. A classification of ontology change. In *Proc. of the 3rd Italian Semantic Web Workshop Scuola Normale Superiore*, Pisa, Italy, 18-20 December 2006.

[FPSS96] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. In Fayyad et al. [FPSSU96], pages 37–54. 625 pages.

[FPSSU96] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors. *Advances in Knowledge Discovery and data Mining*. American Associatation for Artificial Intelligence, 1996. 625 pages.

[FR00] S. Ferre and O. Ridoux. A logical generalization of formal concept analysis. In *Proceedings of the International Conference on Conceptual Structures, ICCS ?2000*, 2000.

[FS07] R. Feldman and J. Sanger. *The Text Mining Handbook*. Cambridge University Press, 2007.

[Gam91] L.T.F Gamut. *Logic Language and Meaning – Intensional Logic and Logical Grammar*, volume 2. The University Of Chicago Press, 1991. A collective pseudonym for J. van Benthem, J. Groenendijk, D. de Jongh, M. Stokhof and H. Verkuyl.

[Gan84] B. Ganter. Two basic algorithms in concept analysis. Technical Report Technical Report 831, Technische Hochschule, Darmstadt, 1984.

[Gar98] D. Garcia. *Analyse automatique des textes pour l'organisation causale des actions, réalisation du système COATIS*. PhD thesis, Thèse d'informatique, Université Paris IV, 1998.

[GD86] J.-L. Guigues and V. Duquenne. Familles minimales d'implication informatives résultant d'un tableau de données binaires. *Mathématiques, Informatique et Sciences Humaines*, 95(24) :5–18, 1986.

[GH07a] L. Geng and H. J. Hamilton. Choosing the right lens : Finding what is interesting in data mining. In Guillet and Hamilton [GH07b].

[GH07b] F. Guillet and H. J. Hamilton, editors. *Quality Measures in Data Mining*, volume 43 of *Studies in Computational Intelligence*. Springer, 2007.

[GK01] B. Ganter and S.O. Kuznetsov. Pattern structures and their projections. In H.S. Delugach and G. Stumme, editors, *Conceptual Structures : Broadening the Base, Proceedings of the 9th International Conference on Conceptual Structures, ICCS 2001, Stanford, CA*, Lecture Notes in Computer Science 2120, pages 129–142. Springer, 2001.

[GLL05] C. Grover, A. Lascarides, and M. Lapata. A comparison of parsing technologies for the biomedical domain. *Natural Language Engineering*, 11(1) :27–65, 2005.

[GM94] R. Godin and R. Missaoui. An incremental concept formation approach for learning from databases. *Theoretical Computer Science*, 133 :378–419, 1994.

[GP04] A. Gómez-Pérez. Ontology evaluation. In *Handbook on Ontologies in information systems*, International Handbooks on Information Systems, chapter 13, pages 251–274. Springer, 2004.

[GPFLC04] A. Gómez-Pérez, M. Fernándo-Lòpez, and O. Corcho. *Ontological Engineering*. Springer, 2004.

[Gui00] S. Guillaume. *Traitement des données volumineuses : Mesures et algorithmes d'extraction de règles d'association et règles ordinales*. PhD thesis, Université de Nantes, 2000.

[GW99] B. Ganter and R. Wille. *Formal Concept Analysis, Mathematical Foundations*. Springer, 1999.

[GW00a] N. Guarino and C. Welty. Towards a methodology for ontology-based model engineering. In *Proceedings of the ECOOP-2000 Workshop on Model En- gineering*, 2000. Available from http ://www.ladseb.pd.cnr.it/infor/ontology/Papers/OntologyPapers.html.

[GW00b] N. Guarino and C.A. Welty. A formal ontology of properties. In *Knowledge Acquisition, Modeling and Management*, pages 97–112, Juan-les-Pins, France, 2000.

[GW02] N. Guarino and C. Welty. Identity and subsumption. In R. Green, C. A. Bean, and S. H. Myaeng, editors, *The Semantics of Relationships : An Interdisciplinary Perspective*, pages 111–125. Dordrecht : Kluwer, 2002.

[GW04] N. Guarino and C. Welty. An overview of ontoclean. In S. Staab and R. Studer, editors, *The Handbook on Ontologies*, pages 151–172. Springer-Verlag, 2004.

[hab04] B. habegger. *Extraction d'informations à partir du Web*. PhD thesis, Université de Nantes, Discipline : informatique, décembre, 2004.

[Har68] Z. Harris. *Mathematical Structures of Languages*. Wiley-Interscience, New-York, 1968.

[HDG00] K. Humphreys, G. Demetriou, and R. Gaizauskas. Two applications of information extraction to biological science journal articles : enzyme interactions and protein structures. In *Proceedings of the Pacific Symposium on Biocomputing (PSB-2000)*, pages 505–516, Honolulu, Hawaii, USA, January 2000.

[Hea04] M. Hearst. Untangling text data mining. In *The 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, 2004.

[HNRHV07] Marianne Huchard, Amedeo Napoli, Mohamed Rouane-Hacene, and Petko Valtchev. Mining description logics concepts with relational concept analysis. In P. Brito, P. Bertrand, G. Cucumel, and F. De Carvalho, editors, *Selected Contributions in Data Analysis and Classification*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 259–270. Springer, Berlin, 2007.

[HNRV07a] M. Huchard, A. Napoli, M. Hacene Rouane, and P. Valtchev. A proposal for combining formal concept analysis and description logics for mining relational data. In S.O. Kuznetsov and S. Schmidt, editors, *proceeding of the 5th International Conference Formal Concept Analysis (ICFCA'07)*, LNAI 4390, pages 51–65, Clermond-Ferrand, France, 2007. Springer, Berlin.

[HNRV07b] M. Huchard, A. Napoli, M. Hacene Rouane, and P. Valtchev. A proposal for combining formal concept analysis and description logics for mining

relational data. In *proceeding of the 5th International Conference Formal Concept Analysis (ICFCA'07)*, Clermond-Ferrand, France, 2007.

[Jac94] C. Jacquemin. Fastr : A unification-based front-end to automatic indexing. In *Actes de la conférence sur la recherche d'information assistée par ordinateur (RIAO'94)*, pages 34–47, New York, USA, 1994. CID, Paris, FRANCE.

[Jac01] C. Jacquemin. *Spotting and Discovering Terms through NLP*. MIT Press, Cambridge MA, 2001.

[JCK+04] D. Janetzko, H. Cherfi, R. Kennke, A. Napoli, and Y. Toussaint. Knowledge-based Selection of Association Rules for Text Mining. In R. Lopez de Màntaras and L. Saitta, editors, *16h European Conference on Artificial Intelligence - ECAI'04*, pages 485–489, Valencia, Spain, 2004. IOS Press.

[JCSZ10] C. Jiang, F. Coenen, R. Sanderson, and M. Zito. Text classification using graph mining-based feature extraction. *Knowledge-Based Systems*, 23(4) :302–308, 2010.

[Jil09] I. Jilani. *Extraction automatiqu de connaissances à partir de textes biomédicaux*. PhD thesis, Université Paris VI, spécialité informatique biomédicale, 2009.

[JM00] D. Jurafsky and J. H. Martin. *Speech and Language Processing*. Prentice-Hall, 2 edition, 2000.

[Jou93] C. Jouis. *Contribution à la conceptualisation et à la modélisation des connaissances à partir d'une analyse de textes. Réalisation d'un prototype : le système SEEK*. PhD thesis, Thèse d'informatique, EHESS, Paris, 1993.

[JR94] Christian Jacquemin and Jean Royauté. Retrieving terms and their variants in a lexicalised unification-based framework. In *Proceedings of the 17th ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1058–1063, 1994.

[Kam81] H. Kamp. A theory of truth and semantic representation. In J.A.G. Groenendijk, T.M.V. Janssen, , and M.B.J. Stokhof, editors, *Formal Methods in the Study of Language*, pages 277–322. Mathematical Centre Tracts 135, Amsterdam, 1981.

[Kay11] M. Kaytoue. *Traitement de données numériques par analyse formelle de concepts et structures de patrons*. PhD thesis, Université Henri Poincaré – Nancy 1, avril 2011.

[Kle04] M. Klein. *Change Management for Distributed Ontologies*. PhD thesis, Vrije Universiteit Amsterdam, August 2004.

[KMR+94] M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen, and A. I. Verkamo. Finding interesting rules from large sets of discovered association rules. In B. K. Bhargava, N. R. Adam, and Y. Yesha, editors, *Proceedings od the 3rd International Conference on Information and Knowledge Management (CIKM'94)*, pages 401–407, Gaithersburg, USA, 1994. ACM Press.

[KNT05] H. Kou, A. Napoli, and Y. Toussaint. Application of text categorization to astronomy field. In *actes de la conférence Applications of Natural Language to Data Bases (NLDB'05)*, pages 32–43, 2005.

[KO02] S.O. Kuznetsov and S.A. Obiedkov. Comparing performance of algorithms for generating concept lattices. *Journal of Experimental & Theoretical Artificial Intelligence*, 14(2/3) :189–216, 2002.

[Koe06] B. Koester. Conceptual knowledge retrieval with fooca : Improving web search engine results with contexts and concept hierarchies. In P. Perner, editor, *Advances in Data Mining, Applications in Medicine, Web Mining, Marketing, Image and Signal Mining, 6th Industrial Conference on DataMining, ICDM2006*, number 4065 in LNCS, pages 176–190, Leipzig, Germany, 2006.

[Koh84] T. Kohonen. *Self-Organization and Associative Memory*. Springer Verlag, 2 edition, 1984.

[KP05] A. Kao and S. Poteet. Text mining and natural language processing : Introduction for the special issue. *ACM SIGKDD Explorations Newsletter*, 7(1) :1–2, 2005.

[KR05] Y. S. Koh and N. Rountree. Finding sporadic rules using apriori-inverse. In T.B. Ho, D. Cheung, and H. Liu, editors, *Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, number 3518 in LNAI, pages 97–106, 2005.

[KS05] S.O. Kuznetsov and M.V. Samokhin. Learning closed sets of labeled graphs for chemical applications. In S. Kramer and B. Pfahringer, editors, *Proceedings of 15th International Conference on Inductive Logic Programming (ILP 2005)*, LNCS 3625, pages 190–208. Springer, 2005.

[KSS08] P. Knoth, M. Schmidt, and P. Smrz. Information extraction – state-of-the-art. Technical report, Report from the "Knowledge in a Wiki" Project, ICT-2007.4.2-211932 (EU Seventh Framework Programme (FP7)), 2008.

[KSV07] M. Krötzsch, S. Schaffert, and D. Vrande ?i ? Reasoning in semantic wikis. In G. Antoniou, U. Abmann, C. Baroglio, S. Decker, N. Henze, P.-L. Patranjan, and R. Tolksdorf, editors, *Reasoning Web*, volume 4636 of *Lecture Notes in Computer Science*, pages 310–329. Springer Berlin / Heidelberg, 2007.

[KUKND10] M. Kaytoue-Uberall, S.O. Kuznetsov, A. Napoli, and S. Duplessis. Mining Gene Expression Data with Pattern Structures in Formal Concept Analysis. *Information Science*, 2010.

[Kuz04] S.O. Kuznetsov. Machine learning and formal concept analysis. In P.W. Eklund, editor, *Concept Lattices, Second International Conference on Formal Concept Analysis, ICFCA 2004, Sydney, Australia*, Lecture Notes in Computer Science 2961, pages 287–312. Springer, 2004.

[Kuz07] S. Kuznetsov. On stability of a formal concept. *Annals of Mathematics and Artificial Intelligence*, 49(1-4) :101–115, April 2007.

[Kuz09] S.O. Kuznetsov. Pattern structures for analyzing complex data. In H. Sakai, M.K. Chakraborty, A.E. Hassanien, D. Slezak, and W. Zhu, editors, *Proceedings of the 12th International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing*, volume 5908 of *LNCS 5908*, pages 33–44. Springer, 2009.

[KWT04] M. Klopotek, S. Wierzchon, and K. Trojanowski, editors. *Intelligent Information Processing and Web Mining, Proceedings of the International IIS : IIPWM'04 Conference held in Zakopane, Poland, May 17-20, 2004*, Advances in Soft Computing. Springer, 2004.

[Lak72] G. Lakoff. Hedges : A study in meaning criteria and the logic of fuzzy concepts. In P. Peranteau, J. Levi, and G. Phares, editors, *Papers from the Eighth Regional Meeting, Chicago Linguistics Society (CLS 8)*, pages 183–228, 1972.

[LC04] N. Lucas and B. Crémilleux. Fouille de textes hiérarchisée appliquée à la détection de fautes. *Document numérique*, 8(3) :107–134, 2004.

[LFF09] J. Li, A. W. Fu, and P. Fahey. Efficient discovery of risk patterns in medical data. *Artificial Intelligence in Medicine*, 45 :77–89, 2009.

[LFZ99] N. Lavrac, P. Flach, and B Zupan. Rule evaluation measures : A unifying view. In *Proceedings of the 9th International Workshop on Inductive Logic Programming (ILP'99)*, volume 1634 of *Lecture Notes in Artificial Intelligence – LNAI*, pages 174–185, Bled, Slovenia, 1999. Springer Verlag.

[LHC97] B. Liu, W Hsu, and S. Chen. Using general impressions to analyse discovered association rules. In *3rd International Conference on Kowledge Discovery and Data Mining (KDD'97*, pages 31–36, Newport Beach, USA, 1997. AAAI Press.

[LHM98] B. Liu, W Hsu, and Y. Ma. Integrating classification and association rules mining. In *International Conference on Knowledge Discovery and Data Mining (KDD'98)*, pages 80–86, New York, USA, 1998.

[LHM99] B. Liu, W. Hsu, and Y. Ma. Visually aided exploration of intersting association rules. In N. Zhong and L. Zhou, editors, *3rd Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'99)*, volume 1574 of *Lecture Notes in Computer Science, LNCS*, pages 380–389, 1999.

[LHP01] W Li, J. Han, and J. Pei. Cmar : Accurate and efficient classification based on multiple class- association rules. In *IEEE International Conference on Data Mining (ICDM'01)*, San Jose, USA, 2001.

[LMP01] J. Lafferty, A. Mccallum, and F. Pereira. Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *Proceeding of the 18th International Conference on Machine learning (ICML'01)*, pages 282–289, 2001.

[LNST06] J. Lieber, A. Napoli, L. Szathmary, and Y. Toussaint. First elements on knowledge discovery guided by domain knowledge (kddk). In S. Ben Yahia, E. M. Nguifo, and R. Belohlávek, editors, *Proceedings of 4th International*

Conference on Concept Lattices and Their Applications (CLA), number 4923 in LNAI, pages 22–41. Springer-Verlag, 2006.

[LR09] M. Laignelet and F. Rioult. Repérer automatiquement les segments obsolescents à l'aide d'indices sémantiques et discursifs. In *Actes de la Conférence sur le Traitement Automatique de la Langue Naturelle (TALN'09)*, Senlis, 2009.

[LR10] M. Laignelet and F. Rioult. Repérer automatiquement les segments obsolescents à l'aide d'indices sémantiques et discursifs. *Traitement Automatique de la langue (TAL)*, 51(1) :41 – 63, 2010.

[LT00] J.-C. Lamirel and Y. Toussaint. Combining symbolic and numeric techniques for DL contents classification and analysis. In *First DELOS Workshop on Information seeking, searching and querying in Digital Librairies*, 12 2000.

[LT02] J.-C. Lamirel and Y. Toussaint. Association de méthodes symboliques et numériques pour l'analyse du contenu de bases de données. In *13ème Congrès Francophone AFRIF-AFIA de Reconnaissance des Formes et Intelligence Artificielle - RFIA'2002*, Angers, France, 01 2002.

[LT04] S. Lallich and O. Teytaud. Evaluation et validation de l ?intérêt des règles d ?association. *Revue des Nouvelles Technologies de l ?Information (RNTI)*, page 193 ?218, 2004.

[LTAS03] J.-C. Lamirel, Y. Toussaint, and S. Al Shehabi. A Hybrid Classification Method for Database Contents Analysis. In *The 16th International FLAIRS Conference - FLAIRS 2003*, St. Augustine, Florida, 2003.

[LTD+01] J.-C. Lamirel, Y. Toussaint, J. Ducloy, C. Czysz, and C. François. Réseaux neuronaux avancés pour la cartographie de la science et de la technologie : Application à l'analyse des brevets. In *Veille Stratégique Scientifique et Technologique - VSST'2001*, pages 215–229, Barcelone, Espagne, 10 2001.

[LTFP01] J.-C. Lamirel, Y. Toussaint, C. François, and X. Polanco. Using a Multi-SOM approach for Mapping of Science and Technology. In *actes de la 8th International Conference on Scientometrics and Informetrics (ISSI'2001)*, Sydney, Australie, 08 2001.

[LTS03] J.-C. Lamirel, Y. Toussaint, and S. Al Shehabi. A hybrid classification method for database contents analysis. In *The 16th International FLAIRS Conference - FLAIRS 2003, St. Augustine, Florida*, May 2003.

[Lux91] M. Luxenburger. Implications partielles dans un contexte. *Mathématiques, Informatique et Sciences Humaines*, 113(29) :35–55, 1991.

[Mae02] A. Maedche. *Ontology Learning for the Semantic Web*. Springer, 2002.

[MAN09] Y. Ma, L. Audibert, and A. Nazarenko. Ontologies étendues pour l'annotation sémantique. In *Actes de la Conférence Ingénierie des Connaissances (IC'2009)*, 2009.

[Mas02] F. Masseglia. *Algorithmes et application pour l'extraction de motifs séquentiels dans le domaine de la fouille de données : de l'incrémental au temps réel*. PhD thesis, Université de Versailles Saint Quentin en Yvelines, 2002.

[May07] D. Maynard. Benchmarking of annotation tools. Deliverablle d1.2.2.1.3 of the european network knowledge web, 111 pages, University of Sheffield, 2007.

[MB05] R. J. Mooney and R. Bunescu. Mining knowledge from text using information extraction. *ACM SIGKDD Explorations Newsletter*, 7(1) :3–10, 2005.

[MG95] G. Mineau and R. Godin. Automatic structuring of knowledge bases by conceptual clustering. *IEEE transactions on Knowledge and Data Engineering*, 7(5) :824–829, 1995.

[Mik98] A. Mikheev. Feature lattices for maximum entropy modelling. In *Proceedings of COLING-ACL'98*, pages 848–854, 1998.

[MLK07] A. Markov, M. Last, and A. Kandel. Fast categorization of web documents represented by graphs. In *WebKDD*, volume 4811 of *Lecture Notes in Computer Science*, pages 56–71. Springer, 2007.

[MLK08] A. Markov, M. Last, and A. Kandel. The hybrid representation model for web document classification. *Int. J. Intell. Syst.*, 23(6) :654–679, 2008.

[MN03] R. J. Mooney and U. Y. Nahm. Text mining with information extraction. In W. Daelemans, T. du Plessis, C.Snyman, and L. Teck, editors, *Multilingualism and Electronic Language Management : Proceedings of the 4th International MIDP Colloquium*, pages 141–160, Bloemfontein, South Africa, September 2003. Van Schaik Pub.,South Africa, 2005.

[MNSY06] S. Maumus, A. Napoli, L. Szathmary, and Y.Toussaint. Réflexions sur l'extraction de motifs rares. In M. Nadif and F.-X. Jollois, editors, *13ièmes rencontre de la Société Francophone de Classification - SFC-06*, pages 157–162, Metz France, 2006.

[Mon73] R. Montague. The proper treatment of quantification in ordinary english. In P. Portner and B. H. Partee, editors, *Formal Semantics : The Essential Readings*, pages 221–242. Blackwell Publishers Ltd, Oxford, UK, 1973. Published Online : 28 JAN 2008.

[Mor99] E. Morin. Automatic acquisition of semantic relations between terms fromtechnical corpora. In *5th International Congress on terminology and Knowledge engineering, TKE'99*, 1999.

[MPRT97] C. Muller, X. Polanco, J. Royauté, and Y. Toussaint. Acquisition et structuration des connaissances en corpus : éléments méthodologiques. Technical report, Rapport de Recherche INRIA RR3198, 1997. 45 pages.

[MRS08] C. D. Manning, P. Raghavan, and H. Schütze. *An Introduction to Information Retrieval*. Cambridge University Press, 2008.

[MS00a] A. Maedche and S. Staab. Discovering conceptual relations from text. In W.Horn, editor, *Proceedings of the 14th European Conference on Artificial Intelligence (ECAI'00)*, pages 321 – 325, Berlin, 2000. IOS Press, Amsterdam.

[MS00b] A. Maedche and S. Staab. Mining ontologies from text. In R. Dieng and O. Corby, editors, *Proc. of the 12 International Conference on Knowledge*

Engineering and Knowledge Management (EKAW'00), volume 1937 of *Lecture Notes in Artificial Intelligence - LNAI*, pages 189–202, Juan-les-Pins, 2000. Springer-Verlag.

[MS01a] A. Maedche and S. Staab. Comparing ontologies - similarity measures and a comparison study. Technical report, Internal Report 408, Institute AIFB, University of Karlsruhe, 2001.

[MS01b] A. Maedche and S. Staab. Ontology learning for the semantic web. *IEEE Intelligent Systems, Special Issue on Semantic Web*, 16(1) :72–79, march 2001.

[MS03] A. Maedche and S. Staab. Ontology learning. In S. Staab and R. Studer, editors, *Handbook on Ontologies in Information Systems Learning*. Springer, 2003.

[MTP04] F. Masseglia, M. Tesseire, and P. Poncelet. Extraction de motifs séquentiels, problèmes et méthodes. *ingénierie des Systèmes d'Information (ISI), numéro spécial sur l'extraction et usages multiples de motifs dans les bases de données*, 9(3-4) :183–210, 2004.

[MTV94] H. Mannila, H. Toivonen, and A. I. Verkamo. Efficient algorithms for discovering association rules. In *Proceeding of the AAAI workshop on Knowledge Discovery in Databases (KDD'94)*, pages 181–192, Seattle, WA, USA, 1994. AAAI Press.

[Nam00] F. Namer. Flemm : Un analyseur flexionnel du français à base de règles. In C. Jacquemin, editor, *TAL, numéro spécial "Traitement automatique des langues pour la recherche d'information"*. Hermes, 2000.

[Nam04] F. Namer. Automatiser l'analyse morpho-sémantique non affixale : le système dérif. *Cahiers de Grammaire*, 28(3), 2004.

[Ned04] C. Nedellec. *Text Mining and its Applications*, chapter Machine learning for information Extraction in Genomics – state of the art and perspectives, pages 95–115. Studies in Fuzziness and Soft Computing. Springer, s. sirmakessis edition, 2004.

[NFM00] N. F. Noy, R. W. Fergerson, and M. A. Musen. The knowledge model of protégé-2000 : Combining interoperability and flexibility. In *EKAW*, pages 17–32, 2000.

[NM00] U.Y. Nahm and R.J. Mooney. Using information extraction to aid the discovery of prediction rules from texts. In *Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining (KDD'00), workshop on Text Mining*, pages 51–58, Boston, MA, 2000.

[NM01] N. F. Noy and D. L. McGuinness. Ontology development 101 : A guide to creating your first ontology. Technical report, Stanford Knowledge System laboratory, Stanford University, USA, 2001.

[NN05] C. Nédellec and A. Nazarenko. Ontology and information extraction : a necessary symbiosis. In Buitelaar et al. [BCM05], pages 155–171.

[NRD+05] E. Nauer, A. Richard, S. Derriere, F. Genova, A. Napoli, and Y. Toussaint. Construction d'une ontologie de descripteurs UCD en astronomie. Technical report, Rapport de Recherche INRIA, 2005.

[NRD+06] E. Nauer, A. Richard, S. Derriere, F. Genova, A. Napoli, and Y. Toussaint. Construction d'une ontologie de descripteurs UCD en astronomie. In *17e journées francophones d'Ingénierie des connaissances - IC 2006*, Nantes France, 2006.

[NT07] E. Nauer and Y. Toussaint. Dynamical modification of context for an iterative and interactive information retrieval process on the web. In *Fifth International Conference on Concept Lattices and Their Applications - CLA 2007*, Montpellier France, 2007.

[NT08a] E. Nauer and Y. Toussaint. Classification dynamique par treillis de concepts pour la recherche d'information sur le web. In *5ème Conférence de recherche en information et applications - CORIA 2008*, pages 71–86, Trégastel France, 2008.

[NT08b] E. Nauer and Y. Toussaint. CreChainDo : an iterative and interactive Web information retrieval system based on lattices. *International Journal of General Systems*, 2008.

[NT08c] Emmanuel Nauer and Yannick Toussaint. Classification dynamique par treillis de concepts pour la recherche d'information sur le web. In *5ème Conférence de recherche en information et applications (CORIA'2008)*, Trégastel, France, 2008.

[NV06] R. Navigli and P. Velardi. Ontology enrichment through automatic semantic annotation of on-line glossaries. In *15th International Conference in Knowlegde Engineering and knowledge Management (EKAW 2006)*, pages 126–140, Czech Republic, 2006. Springer.

[NVB01] C. Nédellec, M. Ould Abdel Vetah, and P. Bessières. Sentence filtering for information extraction in genomics, a classification problem. In *actes de Conference on Practical Knowledge Discovery in Databases, PKDD'2001*, pages 326–338, Freiburg, septembre 2001.

[OHTT01] T. Ono, H. Hishigaki, A. Tanigami, and T. Takagi. Automated extraction of information on protein-protein interactions from the biological literature. *Bioinformatics*, 2001.

[PBTL99a] N. Pasquier, Y. Bastide, R. Taouil, and L.Lakhal. Discovering frequent itemsets for association rules. In *Lecture Notes in Computer Science*, volume 1540, pages 398–416. Springer Verlag, 1999.

[PBTL99b] N. Pasquier, Y. Bastide, R. Taouil, and L.Lakhal. Pruning close itemset lattices for mining association rules. *International Journal of information Systems*, 24(1) :25–46, 1999.

[PC09] M. Plantevit and T. Charnois. Motifs séquentiels pour l'extraction d'information : illustration sur le problème de la détection d'interactions entre gènes. In *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN'09)*, Senlis, France, 2009.

[PCK+09] M. Plantevit, T. Charnois, J. Kléma, C. Rigotti, and B. Crémilleux. Combining sequence and itemset mining to discover named entities in biomedical texts : a new type of pattern. *International Journal of Data Mining, Modelling and Management*, 1(2) :119–148, 2009. Inderscience Publishers.

[PCZ+02] J. Pustejovsky, J. Castaño, J. Zhang, M. Kotecki, and B. Cochran. Robust relational parsing over biomedical literature : Extracting inhibit relations. In *Pacific Symposium on Biocomputing (PSB'02)*, pages 362–373, Kauai, Hawaii, 2002.

[PM04] F. Peng and A. McCallum. Accurate information extraction from research papers using conditional random fields. In *Proceedings of Human Language Technology Conference / North American Association for Computational Linguistics Annual Meeting (HLT-NAACL-2004)*, pages 329–336, Boston, MA, 2004.

[PNV+10] F. Pennerath, G. Niel, P. Vismara, P. Jauffret, C. Laurenço, and A. Napoli. A graph-mining method for the evaluation of bond formability. *ACS Journal of Chemical Information and Modeling*, 50(2) :221–239, 2010.

[Poi03] T. Poibeau. *Extraction automatique d'information. Du texte brut au web sémantique.* Hermès, Paris, 2003.

[PS04] B. Parsia and E. Sirin. Pellet : An owl dl reasoner. In *3rd International Semantic Web Conference (ISWC2004)*, 2004.

[PSF91] G. Piatetsky-Shapiro and W. Frawley. *Knowledge Discovery in Databases.* AAAI Press, 1991.

[Qui90] J. R. Quinlan. Learning logical definitions from relations. *Machine Learning*, 5 :239 ?266, 1990.

[Rab02] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2) :257–286, August 2002.

[RAOTK04] M. Roche, J. Azé, O.Matte-Tailliez, and Y. Kodratoff. Mining texts by association rules discovery in a technical corpus. In Klopotek et al. [KWT04], pages 89–98.

[RH08] M. Rouane-Hacène. *Analyse Relationnelle de Concepts.* PhD thesis, Université de Montréal, 2008.

[RHNV+08a] M. Rouane-Hacene, A. Napoli, P. Valtchev, Y. Toussaint, and R. Bendaoud. Ontology learning from text using relational concept analysis. In P. Kropf, M. Benyoucef, and H. Mili, editors, *International Conference on eTechnologies (MCETECH 08), Montréal*, pages 154–163. IEEE Computer Society, 2008.

[RHNV+08b] M. Rouane Hacene, A. Napoli, P. Valtchev, Y. Toussaint, and R. Bendaoud. Ontology Learning from Text using Relational Concept Analysis. In *International MCETECH Conference on e-Technologies - MCETECH 2008*, Montréal Canada, 2008.

[RHTV09] M. Rouane-Hacene, Y. Toussaint, and P. Valtchev. Mining safety signals in spontaneous reports database using concept analysis. In *Actes de la conférence 12th Conference on Artificial Intelligence in Medicine (AIME'09)*, page 10 pages, 2009.

[Ril93] E. Riloff. Automatically constructing a dictionary for information extraction tasks. In *actes de la 11th National Conference on Artificial Intelligence*, pages 811–816, Washington, DC, USA, Juillet 1993. The AAAI Press/The MIT Press.

[Ril98] E. Riloff. The sundance sentence analyzer, 1998. http ://www.cs.utah.edu/pro jects/nlp/.

[Roc04] M. Roche. *Intégration de la construction de la terminologie de domaines spécialisés dans un processus global de fouille de textes*. PhD thesis, Université Paris-Sud (Orsay), 13 décembre 2004.

[RT96] J. Royauté and Y. Toussaint. Analysing information from large documentary bases. *ERCIM News, Special Issue on Computational Linguistics*, 26, 1996.

[RT97] J. Royauté and Y. Toussaint. Analyse linguistique informétrique pour l'acquisition et la structuration de connaissances. In *Journées Terminologie et Intelligence Artificielle (TIA)*, Toulouse, avril 1997.

[SA95] R. Srikant and R. Agrawal. Mining generalized association rules. In *Proc. of the 21st Int'l Conference on Very Large Databases*, Zurich, Switzerland, 1995.

[SAA+99] G. Schreiber, H. Akkermans, A. Anjewierden, R. DEhoog, N. Shadbolt, W. Vandevelde, and B. Wielinga. *Knowledge engineering and management : the Common-Kads Methodology*. MIT Press, 1999.

[Sah99] S. Sahar. Interestingness via what is not interesting. In S. Chaudhuri et D. Madigan, editor, *Proc. of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'99)*, pages 332–336, San Diego, USA, 1999. ACM Press.

[Sal89] G. Salton. *Automatic Text Processing : The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley, Reading, MA, 1989.

[SBLK05] A. Schenke, H. Bunke, M. Last, and A. Kandel, editors. *Graph-Theoretic Techniques for Web Content Mining*. Machine Perception and Artificial Intelligence 62. World Scientific Publishing, Singapore, 2005.

[SC05] S. Sarawagi and W. W. Cohen. Semi-markov conditional random fields for information extraction. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 1185–1192. MIT Press, Cambridge, MA, 2005.

[SCAG+10] S. Szulman, J. Charlet, N. Aussenac-Gilles, A. Nazarenko, V. Teguiak, and E. Sardet. Dafoe : an ontology building platform from texts or thesauri.

In *Proc. of International Conference on Knowledge Engineering and Knowledge Management (EKAW 2010)*, 2010.

[Sch94] H. Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49, 1994.

[SDT89a] P. Saint-Dizier and Y. Toussaint. Une modélisation logique des relations grammaticales et de la sémantique lexicale pour l'analyse du langage naturel. In *actes de la conférence Reconnaissance des Formes et Intelligence Artificielle, RFIA*, volume 3, pages 1407–1416, Paris, novembre 1989.

[SDT89b] Patrick Saint-Dizier and Yannick Toussaint. Types, essential and functional relations in logic programming for natural language processing. In *Actes du Workshop on Prolog as the Implementation language for Natural Language Processing*, Stockholm, Avril 1989. SICS.

[Seb02] P. Sebillot. Apprentissage sur corpus de relations lexicales sémantiques - la linguistique et l'apprentissage au service d'applications du traitement automatique des langues, 2002. Habilitation à diriger des recherches, Université de Rennes 1.

[Ser08] B. Sertkaya. *Formal Concept Analysis Methods for Descriptions Logics*. PhD thesis, Dresden university, 2008.

[SM91] D. Skuce and I. Meyer. Terminology and knowledge engineering : Exploring a symbiotic relationship. In *6th International Workshop on Knowledge Acquisition for Knowledge-based Systems*, 1991.

[SM01a] G. Stumme and A. Maedche. Fca-merge : Bottom-up merging of ontologies. In *acte de 17th International Joint Conferences on Artificial Intelligence (IJCAI'01)*, pages 225–234,, San Francisco, CA, 2001. Morgan Kaufmann Publishers, Inc.

[SM01b] G. Stumme and A. Maedche. Fca-merge : Bottom-up merging of ontologies. In *17th International Joint Conferences on Artificial Intelligence (IJCAI'01)*, pages 225–234, San Francisco, CA, 2001. Morgan Kaufmann Publishers, Inc.

[Sma93] F. Smadja. Retrieving collocations from texts : Xtract. *Computational Linguistics*, 19(1) :143–177, 1993. Special Issue on Using Large Corpora.

[Smi08] B. Smith. Ontology (science). In C. Eschenbach and M. Grüninger, editors, *Proceedings of the Fifth International Conference on Formal Ontology in Information Systems (FOIS'2008)*, pages 21 – 35. IOS Press, 2008.

[SMP+06] L. Szathmary, S. Maumus, P. Pierre, Y. Toussaint, and A. Napoli. Vers l'extraction de motifs rares. In Chabane Djeraba Gilbert Ritschard, editor, *Extraction et gestion des connaissances*, volume 2/RNTI-E-6, pages 499–510, Lille France, 2006. Cépaduès-éditions.

[SN06] L. Szathmary and A. Napoli. CORON : A Platform for Itemset Mining Algorithms. In *Fourth International Conference on Formal Concept Analysis – ICFCA '06, Dresden, Germany*, Feb 2006. (oral communication and demo, without paper).

[SNV07] L. Szathmary, A. Napoli, and P. Valtchev. Towards rare itemset mining. In *Proceedings of 19th IEEE Intl. Conf. on Tools with Artificial Intelligence (ICTAI '07)*, pages 305–312. IEEE Computer Society, 2007.

[Sou07] A. Soulet. Résumer les contrastes par l'extraction récursive de motifs. In *Actes de la Conférence Francophone sur l'apprentissage automatique (CAP'07)*, 2007.

[Sow84] J. F. Sowa. *Conceptual Structures :Information Processing in Mind and machine*. Addison-Wesley, 1984.

[Sow08] J. F. Sowa. Conceptual graphs. In F. Van Harmelen, V. Lifschitz, and B. Porter, editors, *Handbook in Knowledge Representation*, chapter 5, pages 213–237. Elsevier, 2008.

[STNB11] L. Shi, Y. Toussaint, A. Napoli, and A. Blansché. Mining for reengineering : an application to semantic wikis using formal and relational concept analysis. In *European Semantic Web Conference (ESWC'2011*, pages 421–435, 2011. to be published.

[Sto04] L. Stojanovic. *Methods and Tools for Ontology Evolution*. PhD thesis, Karlsruhe University, 2004.

[SVNG08] L. Szathmary, P. Valtchev, A. Napoli, and R. Godin. Constructing iceberg lattices from frequent closures using generators. In J.-F. Boulicaut, M. R. Berthold, and T. Horv´ath, editors, *Discovery Science (DS'2008)*, number 5255 in LNCS (LNAI), pages 136–147. Springer, Heidelberg, 2008.

[Sza06] L. Szathmary. *Symbolic Data Mining Methods with the Coron Platform*. PhD Thesis in Computer Science, Université Henri Poincaré – Nancy 1, France, Nov 2006.

[TB93] Yannick Toussaint and Mario Borillo. Ingénierie linguistique et spécification de systèmes. In *actes de la conférence Integrated Logistics and Concurrent Engineering (ILCE'93)*, pages 249–258, mars 1993.

[TB94] Yannick Toussaint and Mario Borillo. *Encyclopedia of Software Engineering*, volume 1, chapter Natural Language in Software Engineering, pages 609–613. J. Wiley & Son Inc., 1994.

[TBB91a] Y. Toussaint, M. Borillo, and A. Borillo. Linguistic engineering for software design in the space domain. In European Space Agency, editor, *Actes du workshop on Artificial Intelligence and Knowledge Based Systems for Space*, Mai 1991.

[TBB+91b] Y. Toussaint, M. Borillo, A. Borillo, N. castell, and D. Latour. Applying linguistic engineering to software engineering. In *Actes du 26ème Colloque de Linguistiques*, Poznan, Pologne, Septembre 1991.

[TC02] K. Takeuchi and N. Collier. Use of support vector machines in extended named entity recognition. In *Proceedings of the Sixth Conference on Natural Language Learning (CoNLL-2002)*, volume 20, Taipei, Taiwan, 2002.

[TLD01] Y. Toussaint, J.-C. Lamirel, and M. D'Aquin. Combining Symbolic and Numeric Techniques for Database Content Analysis. In *AI/IEA01*, 2001.

[TND+98] Y. Toussaint, F. Namer, B. Daille, C. Jacquemin, J. Royauté, and N. Hathout. Une approche linguistique et statistique pour l'analyse de l'information en corpus. In *Actes de la conférence sur le Traitement Automatique de la Langue (TALN98)*, pages 182–191, Paris, juin 1998.

[TNPT05] S. Tenier, A. Napoli, X. Polanco, and Y. Toussaint. Knowledge extraction from webpages. In *5th International Workshop on Knowledge Markup and Semantic Annotation - SemAnnot 2005*, Galway/Ireland, 11 2005.

[TNPT06] S. Tenier, N. Napoli, X. Polanco, and Y. Toussaint. Annotation sémantique de pages web. In Chabane Djeraba Gilbert Ritschard, editor, *6èmes journées francophones "Extraction et Gestion de Connaissances" - EGC 2006*, volume 1/RNTI-E-6, pages 305–310, ENIC Telecom - USTL - Lille 1 France, 2006. Cépaduès-éditions.

[Tou92a] Y. Toussaint. *Méthodes informatiques et linguistiques pour l'aide à la spécification de logiciels.* PhD thesis, Université Paul Sabatier de Toulouse, 1992.

[Tou92b] Yannick Toussaint. Les spécifications de logicielles et leur structure textuelle. In *Atelier international PRESCOT "Des architectures textuelles à leur traitement Cognitif*, Chateau de Mons, 1992.

[Tou93] Y. Toussaint. Méthodes informatiques et linguistiques pour l'aide à la spécification de logiciels. Séminaire au Laboratoire d'Informatique de Paris Nord (LIPN), avril 1993.

[Tou94] Y. Toussaint. From natural specification to formal specification : linguistic engineering for software specification. In *Workshop on Artificial Intelligence and Software Eng*, Sorrento, Italy, mai 1994.

[Tou95a] Y. Toussaint. Bases terminologiques informatisées. In *Journées Lexicomatiques et Dictionnairiques*, Lyon, septembre 1995.

[Tou95b] Y. Toussaint. De l'analyse sémantique à l'interprétation conceptuelle. In *Journée Terminologie et intelligence Artificielle*, Paris, 1995.

[Tou95c] Y. Toussaint. Structuration des connaissances pour la rédaction de consignes. In *Journée sur les textes de type consigne, Programme de Recherche en Sciences Cognitives de Toulouse*, 1995.

[Tou96a] Y. Toussaint. Articulation texte-diagramme dans la spécification de logiciel. In *Atelier "Le texte procédural : Langage, Action et Cognition"*, Mons, 1996.

[Tou96b] Y. Toussaint. Combining informetrics and linguistics in order to analyse large documentary databases. In K. S. R. Anjaneyulu, M. Sasikumar, and S. Ramani, editors, *Actes de International Conference on Knowledge Based Computer Systems (KBCS'96)*, pages 279–290, Bombay, Inde, 1996. Narosa Pub.

[Tou97] Y. Toussaint. L'analyse de l'information par la construction à partir de textes d'une base de connaissances partielle. Technical report, Journées scientifiques de l'AUPELF-UREF, Lyon, avril 1997.

[Tou03] Y. Toussaint. Le traitement automatique de la langue contre les erreurs judiciaires : une méthodologie d'analyse systématique des textes d'un dossier d'instruction. In *Actes de la conférence sur le Traitement Automatique de la Langue Naturelle (TALN'03)*, pages 403–409, 2003.

[Tou04a] Y. Toussaint. Extraction de connaissances à partir de textes structurés. *Document numérique*, 8 :11–34, 2004.

[Tou04b] Y. Toussaint. Fouille de textes et veille scientifique : le défi de la connaissance. In *Semaine du Document Numérique, SDN'04*, La Rochelle, France, juin 2004.

[Tou06a] Y. Toussaint. Accéder au contenu des textes : l'apport des outils de traitement de la langue et de fouille de textes. In *Pérenniser le document numérique (Séminaire de formation INRIA 2006)*, Ambroise, France, octobre 2006.

[Tou06b] Y. Toussaint. Des outils informatiques pour accéder au contenu des textes : l'apport des outils de traitement de la langue et de fouille de textes. In Bernard Hidoine et Jacques Millet Lisette Calderan, editor, *Pérenniser le document numérique*, pages 83 – 100. ADBS Éditions, 2006.

[Tou06c] Y. Toussaint. Semantic annotation of texts. In *2nd International Workshop on Enterprise and Networked Enterprises Interoperability (ENEI'2006)*, Vienna, Autriche, 2006.

[Tou08] Y. Toussaint. Semantic annotation using background knowledge. In *International Workshop on Advanced Information Systems for Enterprises (IWAISE'08)*, Constantine, Algeria, avril 2008.

[TS00] Y. Toussaint and A. Simon. Building and interpreting term dependencies using association rules extracted from Galois Lattices. In *Recherche d'Informations Assistée par Ordinateur - Content-Based Multimedia Information Access - RIAO'2000*, page 7 p, Paris, France, 2000.

[TSB09] L. Troiano, G. Scibelli, and C. Birtolo. A fast algorithm for mining rare itemsets. In *Ninth International Conference on Intelligent Systems Design and Applications*, pages 1149–1155, 2009.

[TSC00] Y. Toussaint, A. Simon, and H. Cherfi. Apport de la fouille de données textuelles pour l'analyse de l'information. In *Ingénierie des connaissances - IC'2000*, page 10 p, Toulouse, France, 05 2000.

[TT07a] S. Tenier and Y. Toussaint. Classes d'annotation pour l'annotation sémantique. In Nathalie Laublet, Patrick et Aussenac-Gille, editor, *Atelier "Ontologies et Textes" Ontotexte 2007*, pages 49 – 58, Sophia Antipolis France, 2007.

[TT07b] Y. Toussaint and S. Tenier. Annotation sémantique par classification. In Francky Trichet, editor, *Ingénierie des Connaissances*, volume 2, pages 85–96, Grenoble France, 2007. Cépadues éditions.

[TTNP06] S. Tenier, Y. Toussaint, A. Napoli, and X. Polanco. Instantiation of relations for semantic annotation. In Web Intelligence Consortium, editor, *The*

2006 IEEE/WIC/ACM International Conference on Web Intelligence - WI 2006, pages 463–472, Hong Kong Convention and Exhibition Centre/Chine, 12 2006. IEEE Computer Society Press.

[TW02] L. Tanabe and W. J. Wilbur. Tagging gene and protein names in biomedical text. *Bioinformatics*, 18(8) :1124–1132, 2002.

[UCI+06] V. Uren, P. Cimiano, J. Iria, S. Handschuh, M. Vargas-Vera, E. Motta, and F. Ciravegna. Semantic annotation for knowledge management : Requirements and a survey of the state of the art. *Journal of Web Semantics : Science, Services and Agents on the World Wide Web*, 4 :14–28, 2006.

[VM01] P. Valtchev and R. Missaoui. Building galois (concept) lattices from parts : Generalizing the incremental approach. In H. Delugach and G. Stumme, editors, *Proceedings of the ICCS 2001*, volume 2120 of *LNCS*, pages 290–303. Springer Verlag, 2001.

[VMR+07] Matthieu Vernier, Yann Mathet, François Rioult, Thierry Charnois, Stéphane Ferrari, and Dominique Legallois. Classification de textes d'opinions : une approche mixte n-grammes et sémantique. In *Actes du Défi en Fouille de Textes (DEFT'07)*, 2007.

[vR79] K. van Rijsbergen. *Information Retrieval, 2nd ed.* Butterworths, London, 1979.

[VRHM03] P. Valtchev, M. Rouane-Hacene, and R. Missaoui. A generic scheme for the design of efficient on-line algorithms for lattices. In B. Ganter, A. de Moor, and W. Lex, editors, *ICCS 2003*, volume 2746 of *LNCS*, pages 282–295. Springer, Heidelberg, 2003.

[VTLL10] J. Villerd, Y. Toussaint, and A. Lillo-Lelouët. Adverse drug reaction mining in pharmacovigilance data using formal concept analysis. In *Proceedings of the Principles and Practice of Knowledge Discovery in Databases Conference (PKDD'10)*, volume 6323 of *Lecture Notes in Computer Science*, pages 386–401, 2010.

[VVSH08] J. Völker, D. Vrandecic, Y. Sure, and A. Hotho. Aeon - an approach to the automatic evaluation of ontologies. *Journal of Applied Ontology*, 3(1-2) :41–62, 2008.

[WB10] I. Wakwoya-Bayissa. From data to knowledge : Semi-automatic analysis of heterogeneous textual resources over the web. Master's thesis, University of Nancy 2, nancy, France, 2010.

[YH02] X. Yan and J. Han. gspan : Graph-based substructure pattern mining. In *2nd IEEE Int. Conf. on Data Mining (ICDM 2003)*, pages 721 ?–724, Maebashi, Japan, 2002. IEEE Press.

[YTMT01] A. Yakushiji, Y. Tateisi, Y. Miyao, and J.-I. Tsujii. Event extraction from biomedical papers using a full parser. In *Pacific Symposium on Biocomputing*, volume 6, pages 408–419, 2001.

[ZAR03] D. Zelenko, C. Aone, and A. Richardella. Kernel methods for relation extraction. *Journal of Machine Learning Research*, 3(1083–1106), 2003.

[ZH02] M. J. Zaki and C.-J. Hsiao. Charm : An efficient algorithm for closed itemset mining. In *Proceedings of the SIAM International Conference on Data Mining (SDM'02)*, pages 33–43, 2002.

[ZH05] M. J. Zaki and C.-J. Hsiao. Efficient algorithms for mining closed itemsets and their lattice structure. *IEEE Trans. on Knowl. and Data Eng.*, 17(4) :462–478, 205.

[Zim08] A. Zimmermann. *Sémantique des réseaux de connaissances : gestion de l'hétérogénéité fondée sur le principe de médiation.* PhD thesis, Université Joseph Fourier, Grenoble, spécialité informatique, novembre 2008.

[ZPOL97] M. J. Zaki, S. Parthasarathy, M. Ogihara, and W. Li. New algorithms for fast discovery of association rules. In *Proceedings of the 3rd Internationl Conference on Knowledge Discovery in Databases*, pages 283–286, 1997.

Annexe A

Post-traitement des règles d'association [CNT09]

A Conformity Measure using Background Knowledge for Association Rules: Application to Text Mining

Hacène Cherfi
INRIA Sophia Antipolis, 06902 Sophia Antipolis, France
hacene.cherfi@sophia.inria.fr
Amedeo NAPOLI
LORIA - INRIA, 54506 Vandoeuvre-lès-Nancy, France
amedeo.napoli@loria.fr
Yannick TOUSSAINT
LORIA - INRIA, 54506 Vandoeuvre-lès-Nancy, France
yannick.toussaint@loria.fr

ABSTRACT

A text mining process using association rules generates a very large number of rules. According to experts of the domain, most of these rules basically convey a common knowledge, i.e. rules which associate terms that experts may likely relate to each other. In order to focus on the result interpretation and discover new knowledge units, it is necessary to define criteria for classifying the extracted rules. Most of the rule classification methods are based on numerical quality measures. In this chapter, we introduce two classification methods: The first one is based on a classical numerical approach, i.e. using quality measures, and the other one is based on domain knowledge. We propose the second original approach in order to classify association rules according to qualitative criteria using domain model as background knowledge. Hence, we extend the classical numerical approach in an effort to combine data mining and semantic techniques for post mining and selection of association rules. We mined a corpus of texts in molecular biology and present the results of both approaches, compare them, and give a discussion on the benefits of taking into account a knowledge domain model of the data.

KEYWORDS

Knowledge extraction, Text Mining, Association Rules, Semantic Measure, Statistical Measure, Domain knowledge model, Natural language processing.

1. INTRODUCTION

From the data mining point of view, texts are complex data giving raise to interesting challenges. First, texts may be considered as weakly structured, compared with databases that rely on a predefined schema. Moreover, texts are written in natural language, carrying out implicit knowledge, and ambiguities. Hence, the representation of the content of a text is often only partial and possibly noisy. One solution for handling a text or a collection of texts in a satisfying way is to take advantage of a knowledge model of the domain of the texts, for guiding the extraction of knowledge units from the texts.

In this chapter, we introduce a knowledge-based text mining process (KBTM) relying on the knowledge discovery process (KDD) defined in [Fayyad *et al.*, 1996]. The KBTM process relies on an interactive loop, where the analyst – an expert of the text domain – controls and guides the mining process. The objective of the mining process is to enrich the knowledge model of the text domain, and, in turn, to improve the capability of the knowledge-based text mining process itself.

Following a natural language processing of the texts described in [Cherfi *et al.*, 2006], the text mining process (also denoted by TM in the following) is applied to a binary table Texts × Keyterms, and produces a set of association rules (AR in the following). The set Keyterms includes a set of keyterms giving a kind of summary of the content of each text. The extraction of association rules is carried out thanks to a *frequent itemse*t algorithm (namely the Close algorithm [Pasquier *et al.*, 1999]). Association rules show some advantages, among which the facts that AR are easily understandable and that they highlight regularities existing within the set of texts.

Two text mining approaches based on association rules are studied hereafter. The first approach is based on the use of statistical quality measures for classifying the extracted rules [Cherfi et al., 2006]. A set of five quality measures is introduced, each of them expressing some particular aspects of the texts: e.g. rare keyterms, functional dependencies, or probabilistic correlations between keyterms. One limitation of this approach is due to the numerical characteristics of the classification process, which takes into account the distribution of the keyterms, and ignores the semantics carried by the keyterms. By contrast, a second approach is based on a domain knowledge model of the texts which is used to classify the extracted association rules. The knowledge model is a pair $(K, \sqsubseteq)$ where K is a finite set of keyterms and $\sqsubseteq$ is a specialisation relation (*i.e.*, a partial ordering). Hence, the quality of a rule depends on the conformity of the rule with respect to the knowledge model: a rule is interesting if it includes semantic relations that are not already known in the knowledge model. Thus, the knowledge model is used to guide the interpretation and the classification of the extracted association rules. This KBTM approach is original and relies on a qualitative approach rather than on a more classical approach based on statistical quality measures. Two experiments show that the KBTM approach gives substantial and good quality results, opening new perspectives in the difficult field of text mining. The objective of these experiments is to show how far our proposed Conformity measure is consistent with the text mining task in a specific domain (here molecular biology).

This chapter is organised as follows. Firstly, we introduce the context of association rule extraction for text mining, and we present and discuss an example, based on statistical quality measures. Then, we introduce the principles of the KBTM process. We analyse thanks to an example –the same as in the first part of the chapter – the KBTM process for the so-called simple and complex extracted AR. The following section sets up an experiment and a qualitative analysis based on real-world collection of texts with the help of an analyst. The AR are classified according to the conformity measure, in contrast with five statistical measure classifications. We continue the chapter with a discussion on the benefits of the KBTM approach, and we mention some related work. The chapter ends with a conclusion and draws future work.

2. EXTRACTION OF ASSOCIATION RULES FOR TEXT MINING

2.1. Text processing for data mining preparation

In our experiments, we dealt with a collection of texts (hereafter called corpus) in molecular biology. Basically, we start with a set of bibliographical records characterised by contextual metadata, *e.g.*, title, author(s), date, status (whether published or not), keywords, etc. Hereafter, we explain how we get the keyterms associated with each text.

Extracting Textual Fields in the Sources: A first processing of this collection of records consists in extracting two textual fields, the title and the abstract.

Part-of-speech (POS) Tagging: It is a natural language processing (NLP) technique which associates with each word of the texts a linguistic tag corresponding to its grammatical category (noun, adjective, verb, etc.). A POS-tagger needs a learning phase with a manually tagged vocabulary. A POS-tagger basically uses a statistical model to learn how to predict the category of a word with respect to the preceding word categorisation. Several taggers exist for English and show high performance of correctness [Paroubek, 2007]. For example, sentence (1) extracted from one of our texts gives the tagged sentence (2):

1. Two resistant strains were isolated after four rounds of selection.
2. Two/CD resistant/JJ strains/NNS:pl were/VBD isolated/VBN after/IN four/CD rounds/NNS:pl of/IN selection/NN.

Terminological Indexing: In our experiments, the texts have been processed and represented by a set of keyterms. A keyterm is a noun phrase (*i.e.*, one to many words) of our vocabulary which can be associated with a domain concept of our knowledge model, thus, it ensures the transition from the linguistic to the knowledge level.

Keyterm Identification and Variants: We have used the FASTR [Jacquemin, 1994] terminological extraction system for identifying the keyterms of our vocabulary in the text. It allows us to recognise a keyterm in several variant forms. For example, the expression “transfer of capsular biosynthesis genes” is considered as a variant form of the keyterm “gene transfer” which belongs to the vocabulary. However, all the variants are not acceptable; NLP meta-rules are used to keep the variants preserving the initial sense of the keyterm. The keyterm variants are identified using the meta-rules. A meta-rule is a transformation rule operating on the grammatical description of a keyterm and the linguistically authorised variation of this description. For example, the expression “transfer of genes” is recognised as a variation of the keyterm “gene transfer” (which belongs to the vocabulary) by a *permutation* meta-rule of “gene” and “transfer”. The expression “transfer of capsular biosynthesis genes” is recognised as well by applying an *insertion* meta-rule (of “capsular biosynthesis”). In this way, the NLP keyterm identification contributes to reduce the word dispersion in the description of a text by unifying variants to a single keyterm.

2.2. Association Rules and Statistical Quality measures

Let $T = \{t_1 , t_2, ..., t_m\}$ be a set of m texts and $K = \{k_1 , k_2 , ... , k_n\}$ a set of n keyterms associated with these texts. An association rule is a weighted implication such as $A \rightarrow B$ where $A = \{k_1 , k_2, ..., k_p\}$ (the *body*) and $B = \{k_{p+1} , k_{p+2}, ..., k_q\}$ (the *head*). The rule $A \rightarrow B$ means that if a text contains $\{k_1 , k_2, ..., k_p\}$ then it tends to contain also $\{k_{p+1} , k_{p+2}, ..., k_q\}$ with a probability given by the confidence of the rule. Several algorithms aim at extracting association rules: Apriori [Agrawal *et al.*, 1996] or Close [Pasquier et *al.*, 1999] that will be used hereafter. The support and the confidence are two quality measures related to association rules that are used to reduce the number of the extracted units, hence reducing the complexity of the extraction process. The support of a rule $A \rightarrow B$ measures the number of texts containing both keyterms of A and B. The union of the keyterm sets A and B is denoted by $A \sqcap B$. The support may be normalised by the total number of texts. The confidence of a rule is defined by the ratio between the number of texts containing the keyterms in $A \sqcap B$, and the number of texts containing the keyterms in A. The confidence is seen as the conditional probability $P(B/A)$. The confidence of a rule measures the proportion of examples and counterexamples of the rule. A counterexample states that there exist texts having all the keyterms of A, but not necessarily all the keyterms of B. When the confidence of a rule is 1, the rule is *exact*, otherwise it is *approximate*. Two thresholds are defined, σ_s for the minimum support, and σ_c for the minimum confidence. A rule is valid whenever its support is greater than σ_s and its confidence is greater than σ_c.

Considering a rule such as $A \rightarrow B$, if A and B are frequent keyterm sets (*i.e.*, their support is above the σ_s threshold), then they are shared by a large proportion of texts, and the probabilities $P(A)$, $P(B)$, and $P(A \sqcap B)$ are high (here probability stands for the number of texts containing a given keyterm set out of the total number of the texts). The importance of such frequent keyterm sets is rather small, from the KDD point of view. By contrast, when A and B are rare, i.e. they have a low probability, then these keyterm sets are shared by a low number of texts, i.e. the keyterms in A and B may be related in the context of the mined text set. However, the support and the confidence are not always sufficient for classifying extracted association rules in a meaningful way. This reason leads to introduce a number of other statistical quality measures attached to the rules enlightening some particular aspects on the rules [Lavrac *et al.,*1999]. Five of these quality measures are presented hereafter, and have been used in our two experiments.

1. The *interest* measures the degree of independence of the keyterm sets A and B, and is defined by $\text{interest}(A \rightarrow B) = P(A \sqcap B)/P(A) \times P(B)$. The interest is symmetrical ($\text{interest}(A \rightarrow B = \text{interest}(A \rightarrow B)$) and has its range in the interval $[0, +\infty[$. It is equal to 1 whenever the "events" A and B are statistically independent. The more A and B are incompatible, the more $P(A \sqcap B)$, and hence the interest, tend to 0;
2. The *conviction* allows us to select among the rules $A \rightarrow B$ and $A \rightarrow B$ the one having the less counterexamples. The conviction is defined by $\text{conviction}(A \rightarrow B) = P(A) \times P(\neg B)/P(A \sqcap \neg B)$. The conviction is not symmetrical, and has its range in $[0, +\infty[$. It denotes a dependency between A and B whenever it is greater than 1, independence whenever it is equal to 1, and no dependency at all whenever it is lower than 1. The conviction is not computable for exact rules because $P(A \sqcap \neg B)$ is equal to 0 (there is no counterexample for exact rules);
3. The *dependency* measures the distance between the confidence of the rule and the independence case: $\text{dependency}(A \rightarrow B) = |P(B/A) - P(B)|$. This measure has its range in $[0, 1[$, where a

dependency close to 0 (respectively to 1) means that A and B are independent (respectively dependent);

4. The *novelty* is defined by novelty(A → B) = P(A ⊓ B) − P(A) × P(B), and has its range within]−1, 1[, with a negative value whenever P(A ⊓ B) < P(A) × P(B). The novelty tends to −1 for rules with a low support, i.e. P(A ⊓ B) ≃ 0. The novelty is symmetrical although the rule A → B may have more counterexamples than the rule B → A. It leads to the definition of the following measure;
5. The *satisfaction* measure is defined by satisfaction(A → B) = P(¬B)− P(¬B|A)/P(¬B) . The satisfaction has its range in [−∞, 1], and is equal to 0 whenever A and B are independent. The satisfaction cannot be used for classifying exact rules because, in this case, its value is equal to 1.

2.3. Using Quality Measures on a Small Example

An example borrowed from [Pasquier *et al.*, 1999] will be used to illustrate the behaviour of the statistical quality measures introduced above. Let us consider six texts {t_1, t_2, t_3, t_4, t_5, t_6} described by a set of five keyterms, namely {a, b, c, d, e}. So the text t1 is described by the keyterm set {b, c, e} (see Table 1), and hereafter more simply denoted by the bce. The extraction of the association rules has been performed with the Close algorithm [Pasquier *et al.*, 1999]. Twenty association rules, numbered r_1 , . . . , r_{20} , have their support greater than the threshold σ_s = 1/6 (where 6 is the total number of texts), and their confidence is greater than σ_c = 0.1 (or 10%). The set of extracted association rules is given in Table 2. The rules have been extracted from closed frequent keyterm sets. The Close algorithm is based on levelwise search of *closed frequent* keyterm sets in the binary table Texts × Keyterms, starting from the smallest closed keyterm sets {ac, be} to the largest closed keyterm set abce. A closed frequent keyterm set corresponds to a maximal set of keyterms shared by a given subset of texts, with a support greater than the σ_s threshold. Once the closed frequent keyterm sets have been extracted, the association rules of the form $P_2 \rightarrow P_1 \setminus P_2$ may be derived, where for example b → ce stands for "b → bce \ b". The extracted association rules A → B have a minimal *body*, i.e. A corresponds to a generator, and a maximal head, i.e. B corresponds to a closed set for the Galois connection associated with the relation Texts × Keyterms (see for example [Bastide *et al.*, 2000]). For example, the association rules b → e and b → c ∧ e are extracted, because the corresponding keyterm sets be and bce are closed sets in the Galois connection.

Table 1.The textual database

Texts	Keyterms
t_1	acd
t_2	bce
t_3	abce
t_4	be
t_5	abce
t_6	bce

Table 2.The set of 20 valid AR

id	Rule	id	Rule
r_1	$b \rightarrow e$	r_{11}	$a \rightarrow c$
r_2	$b \rightarrow c \wedge e$	r_{12}	$b \wedge c \rightarrow a \wedge e$
r_3	$a \wedge b \rightarrow c \wedge e$	r_{13}	$d \rightarrow a \wedge c$
r_4	$a \rightarrow b \wedge c \wedge e$	r_{14}	$c \rightarrow b \wedge e$
r_5	$b \wedge c \rightarrow e$	r_{15}	$c \rightarrow a \wedge d$
r_6	$b \rightarrow a \wedge c \wedge e$	r_{16}	$c \rightarrow a \wedge b \wedge e$
r_7	$e \rightarrow b \wedge c$	r_{17}	$c \wedge e \rightarrow b$
r_8	$a \wedge e \rightarrow b \wedge c$	r_{18}	$c \wedge e \rightarrow a \wedge b$
r_9	$a \rightarrow c \wedge d$	r_{19}	$e \rightarrow b$
r_{10}	$e \rightarrow a \wedge b \wedge c$	r_{20}	$c \rightarrow a$

The classification of the rules according to the different quality measures is given in Table 3. In each column of the table, the rules are classified according to the value of the measure in a decreasing order. Such a rule classification may be presented to an analyst, either for the whole set of measures or only one particular measure. An algorithm for classifying extracted association rules according to these quality measures (and their roles) is proposed in [Cherfi *et al.*, 2006].

Table3.Statistical measures for the 20 valid AR in a decreasing order

id	support	id	confidence	id	interest	id	conviction	id	dependence	id	novelty	id	satisfaction
r_1	5	r_1	1.000	r_9	2.000	r_7	1.667	r_{13}	0.500	r_1	0.139	r_1	1.000
r_2	5	r_3	1.000	r_{13}	2.000	r_2	1.667	r_3	0.333	r_{19}	0.139	r_3	1.000
r_6	5	r_5	1.000	r_3	1.500	r_{12}	1.333	r_8	0.333	r_2	0.111	r_5	1.000
r_7	5	r_8	1.000	r_8	1.500	r_{18}	1.333	r_1	0.167	r_3	0.111	r_8	1.000
r_{10}	5	r_{11}	1.000	r_{12}	1.500	r_9	1.250	r_5	0.167	r_5	0.111	r_{11}	1.000
r_{14}	5	r_{13}	1.000	r_{18}	1.500	r_{20}	1.250	r_9	0.167	r_7	0.111	r_{13}	1.000
r_{15}	5	r_{17}	1.000	r_1	1.200	r_6	1.111	r_{11}	0.167	r_8	0.111	r_{17}	1.000
r_{16}	5	r_{19}	1.000	r_2	1.200	r_{10}	1.111	r_{12}	0.167	r_9	0.111	r_{19}	1.000
r_{19}	5	r_2	0.800	r_5	1.200	r_{16}	1.111	r_{17}	0.167	r_{11}	0.111	r_2	0.400
r_{20}	5	r_7	0.800	r_6	1.200	r_{15}	1.042	r_{18}	0.167	r_{12}	0.111	r_7	0.400
r_5	4	r_{14}	0.800	r_7	1.200	r_4	1.000	r_{19}	0.167	r_{13}	0.111	r_{12}	0.250
r_{12}	4	r_4	0.667	r_{10}	1.200	r_{14}	0.833	r_2	0.133	r_{17}	0.111	r_{18}	0.250
r_{17}	4	r_{20}	0.600	r_{11}	1.200	r_1	0.000	r_7	0.133	r_{18}	0.111	r_9	0.200
r_{18}	4	r_{12}	0.500	r_{15}	1.200	r_3	0.000	r_{20}	0.100	r_{20}	0.111	r_{20}	0.200
r_4	3	r_{18}	0.500	r_{16}	1.200	r_5	0.000	r_6	0.067	r_6	0.056	r_6	0.100
r_9	3	r_6	0.400	r_{17}	1.200	r_8	0.000	r_{10}	0.067	r_{10}	0.056	r_{10}	0.100
r_{11}	3	r_{10}	0.400	r_{19}	1.200	r_{11}	0.000	r_{16}	0.067	r_{16}	0.056	r_{16}	0.100
r_3	2	r_{16}	0.400	r_{20}	1.200	r_{13}	0.000	r_{14}	0.033	r_{15}	0.028	r_{15}	0.040
r_8	2	r_9	0.333	r_4	1.000	r_{17}	0.000	r_{15}	0.033	r_4	0.000	r_4	0.000
r_{13}	1	r_{15}	0.200	r_{14}	0.960	r_{19}	0.000	r_4	0.000	r_{14}	-0.028	r_{14}	-0.200

3. CONFORMITY OF AN ASSOCIATION RULE WITH RESPECT TO A KNOWLEDGE MODEL

3.1. Conformity for a Simple Rule

Definition 1 (Knowledge Model)

A knowledge model, denoted by $(K, \sqsubseteq)$, is a finite, directed graph with K standing for the set of vertices (the keyterms), and the relation $\sqsubseteq$ defining the edges of the graph and the partial ordering over the keyterms in K. For each $x, y \in K$, $x \sqsubseteq y$ means that each instance of the keyterm concept x is also an instance of the keyterm concept y.

The principle of classifying AR according to their conformity with a knowledge model is stated as follows: we assign a high value of conformity to any association rule $A \rightarrow B$ that is "represented" in $(K, \sqsubseteq)$ with a relation $A \sqsubseteq B$ existing between the keyterms $a_i \in A$ and $b_j \in B$, $i, j \geq 1$. We suppose in the following of this section that the rules are simple in the sense that their *body* and *head* are restricted to a single keyterm, for example $b \rightarrow e$. The so-called complex rules where the *body* and/or the *head* are composed of more than one keyterm are considered in section 3.4.

Definition 2 (Conformity for a Simple AR with the Knowledge Model)

Let k_1, k_2 be in K, and let $k_1 \rightarrow k_2$ be a valid AR. The conformity measure of $k_1 \rightarrow k_2$ with $(K, \sqsubseteq)$ is defined by the probability of finding out a path from k_1 to k_2 – called hereafter the probability transition from k_1 to k_2 – in the directed graph of $(K, \sqsubseteq)$. This path can be composed of one to several edges.

If we consider that updating the knowledge model consists in introducing new keyterms and new relations between keyterms in K, then an association rule $x \rightarrow y$ is conform to $(K, \sqsubseteq)$ (i.e., it has a high value of conformity) if the relation $x \sqsubseteq y$ exists in $(K, \sqsubseteq)$. Otherwise, the rule is not conform to the knowledge model (i.e., its conformity value is low). Indeed, we have to notice that a rule $x \rightarrow y$ extracted within the text mining process is not added to $(K, \sqsubseteq)$ without the control of the analyst in charge of updating a knowledge model of his domain. Any knowledge unit update is supervised by the analyst. The computation of the conformity is based on the principles of the spreading activation theory [Collins & Loftus,1975] stating that the propagation of an information marker through the graph of the knowledge model from a given vertex, say k_1, to another vertex, say k_2, relies on the strength associated to the marker. The value of the strength depends on: (i) the length of the path, and (ii) on the number of reachable keyterms starting from k_1 in $(K, \sqsubseteq)$. The strength of the marker monotonically decreases with respect to these two factors.

Definition 3 (Calculation of the Conformity Measure for Simple Rules)

The conformity of a simple rule $k_1 \rightarrow k_2$ is defined as the transition probability from the keyterm k_1 to the keyterm k_2, and is dependent on the minimal path length between k_1 and k_2, and the centrality of k_1 in $(K, \sqsubseteq)$ which depends on how many keyterms are related to k_1 in $(K \setminus k_1)$.

3.2. Transition Probability

Given the domain knowledge model $(K, \sqsubseteq)$, a probability transition table is set and used as a basis of the conformity calculation. The probability transition of k_i and k_j depends on the minimal distance $d(k_i, k_j)$ between a keyterm k_i and a keyterm k_j in $(K, \sqsubseteq)$. We distinguish two particular cases:

1. For each k_i, $d(k_i,k_i) = 1$ in order to take into account the reflexivity of the relation $\sqsubseteq$, and to avoid abnormally high probabilities in a case where there is no outgoing edge from k_i (as illustrated by the vertex c in Figure 1);
2. If it does not exist a path from a keyterm k_i to a keyterm k_j , then we set a "minimal" (non zero) transition probability by using $d(k_i,k_j) = 2N+1$, where N is the cardinal of the set of keyterms in K.

The transition probability from k_i to k_j , denoted by $Cty(k_i,k_j)$, defines the Conformity measure of the rule $k_i \rightarrow k_j$, and relies on the product of two elements: (i) the distance from k_i to k_j , and (ii) a normalisation factor, denoted by $\delta(k_i)$. Moreover, two additional principles are used:

1. The higher the distance between two keyterms k_i and k_j is, the lower the conformity for $k_i \rightarrow k_j$ is;
2. The normalisation factor of a keyterm k_i depends on all the keyterms in K, either they are reachable from k_i or not. Putting things altogether, the formula for calculating the conformity for a simple rule is stated as follows: $\boxed{Cty(k_i,k_j) = [d(k_i,k_j) \times \delta(k_i)]^{-1}}$ where the normalisation factor of k_i is: $\delta(k_i) = \sum_{x \in K} 1/d(k_i,x)$.

Hence, $\delta(k_i)$ depends on the number of outgoing edges from k_i in K: the higher the number of outgoing edges from k_i is, the lower $\delta(k_i)$ is. In accordance, when there is no outgoing edge from a keyterm k_i; this keyterm k_i becomes "predominant" because the highest transition probability for k_i is the reflexive transition as $d(k_i,k_i) = 1$. The normalisation factor $\delta(k_i)$ is computed only once for each keyterm k_i of the knowledge model, and the following equation holds: $\sum_{x \in K} Cty(k_i,x) = 1$.

3.3. A Small Example for Simple AR

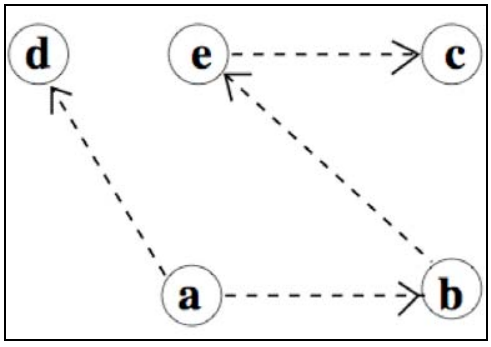


Figure 1. The knowledge model K.

Let Figure 1 be an example of a knowledge model, where an edge between k_i and k_j vertices is interpreted as the specialisation relation $k_i \sqsubseteq k_j$. Based on this model, we may compute the conformity related to each transition as shown in Table 4. Next, we provide details for the computation of the conformity measure for two examples: firstly between a and c where there exists a path in the model, and secondly between c and d, where a path is missing in $(K, \sqsubseteq)$.

$$\begin{aligned} Cty(a, c) &= [d(a,c) \times \Sigma_{x\in\{a,b,c,d,e\}} 1/d(a,x)]^{-1} \\ &= [d(a,c) \times (1/d(a,a) + 1/d(a,b) + 1/d(a,c) + 1/d(a,d) + 1/d(a,e))]^{-1} \\ &= [3 \times (1 + 1 + 1/3 + 1 + 1/2)]^{-1} = 2/23 = 0.09 \end{aligned}$$

$$\begin{aligned} Cty(c, d) &= [d(c,d) \times \Sigma_{x\in\{a,b,c,d,e\}} 1/d(c,x)]^{-1} \\ &= [d(c,d) \times (1/d(c,a) + 1/d(c,b) + 1/d(c,c) + 1/d(c,d) + 1/d(c,e))]^{-1} \\ &= [11 \times (1/11 + 1/11 + 1 + 1/11 + 1/11)]^{-1} = 1/15 = 0.07 \end{aligned}$$

Table4.The conformity scores with the model (K, ⊑) of Figure 1

→	a	b	c	d	e	Σ
a	0.26	0.26	0.09	0.26	0.13	1
b	0.03	0.37	0.19	0.03	0.37	1
c	0.07	0.07	0.73	0.07	0.07	1
d	0.07	0.07	0.07	0.73	0.07	1
e	0.04	0.04	0.44	0.04	0.44	1

Once the computation of the Table 4 is completed, the conformity for each simple rule $k_i \rightarrow k_j$ is given by looking up to the corresponding row i and column j of this table. From the previous example given in Table 2: r_1, r_{19} and r_{11}, r_{20} are two pairs of symmetrical simple rules. Hence, the Table 4 gives their conformity:

(r_{19}) : e→b with $Cty(r_{19}) = 0.04$	(r_{20}) : c→a with $Cty(r_{20}) = 0.07$
(r_1) : b→e with $Cty(r_1) = 0.37$	(r_{11}) : a→c with $Cty(r_{11}) = 0.09$

According to the conformity measure – the classification of the rules is presented in the increasing order – the interesting rules have the lowest values in conformity with $(K, \sqsubseteq)$. For the four previous simple rules, the classification is established as follows: $\{r_{19}, r_{20}, r_{11}, r_1\}$.

The rule r_{11} (in 3rd position in the classification), is already known in $(K, \sqsubseteq)$: its conformity is low because the distance between the two vertices (a and c) is the longest one in $(K, \sqsubseteq)$. The first two rules r_{19} and r_{20} possibly could enrich the model under the supervision of the analyst. It should be noticed that these four rules are classified at very different ranks depending on the statistical measures. Likely because we use an extra knowledge source $(K, \sqsubseteq)$, along with the textual database used for the classification in Table 3.

If an analyst studies the rules sequentially, following any statistical measure, he may be overwhelmed by rules which reflect knowledge already known in $(K, \sqsubseteq)$. Moreover, a major knowledge loss occurs when a number of extracted rules containing new pieces of interesting knowledge are classified at the bottom following the statistical classification lists. On the contrary, the classification given by the conformity measure may draw the attention of the analyst on the possible enrichment of the current domain model $(K, \sqsubseteq)$ with interesting extensions and modifications.

3.4. Conformity for Complex Rules

The complex rules have their left and/or right parts composed of more than one keyterm. Three different kinds of rules may be distinguished. The first is called a 1—m rule: $k_1 \rightarrow k_2 \wedge ... \wedge k_{m+1}$ with $m \geq 2$, and it is composed of one keyterm on the left part and its right part has at least two keyterms. The second is called a n—1 rule: $k_1 \wedge ... \wedge k_n \rightarrow k_{n+1}$ with $n \geq 2$ that has its left part composed of at least two keyterms and a right part with a single keyterm. Finally, an n—m rule: $k_1 \wedge ... \wedge k_n \rightarrow k_{n+1} \wedge ... \wedge k_{n+m}$ where both $(n, m) \geq 2$. We generalize the conformity measure for complex rules by examining its definition for the three kinds (respectively, 1—m, n—1, and n—m) AR.

1—m rules. Let us consider the example of a 1—2 rule: $R_1 : x \rightarrow y \wedge z$. Following predicate logic, R1 can be rewritten in: $\neg x \vee (y \wedge z) = (\neg x \vee y) \wedge (\neg x \vee z)$. This rule can be normalised in a clausal form and decomposed into a conjunction of simple rules: $R_1 = (x \rightarrow y) \wedge (x \rightarrow z)$. Accordingly, the rule R1 is in conformity with $(K, \sqsubseteq)$ if each simple rule of the decomposition is in conformity with $(K, \sqsubseteq)$. The conformity for R1 is then defined by:

$$Cty(R_1 : x \rightarrow y \wedge z) = \min(Cty(x \rightarrow y), Cty(x \rightarrow z))$$

The conformity measure range stands in $[0,1[$. The min function ensures that if at least one simple rule has a low conformity measure, then the complex rule has also a low conformity measure, i.e., the rule may contain some new information for updating $(K, \sqsubseteq)$. Conversely, if all the simple rules have a high conformity measures, i.e., if they all are conform to the model $(K, \sqsubseteq)$, then R_1 is also considered to be conform to $(K, \sqsubseteq)$.

n—1 rules. Let us consider the example of the 2—1 rule R_2: $x \wedge y \rightarrow z$. Following predicate logic, R_2 can be rewritten in: $\neg(x \wedge y) \vee z = (\neg x \vee \neg y) \vee z = (\neg x \vee y) \vee (\neg y \vee z)$. This rule can be decomposed into a disjunction of two simple rules: $R_2 = (x \rightarrow z) \vee (y \rightarrow z)$. Thus, the rule R_2 is in conformity with $(K, \sqsubseteq)$ if one of the simple rules of the decomposition is in conformity with $(K, \sqsubseteq)$. The conformity for R_2 is then defined by: $Cty(R_2 : x \wedge y \rightarrow z) = \max(Cty(x \rightarrow z), Cty(y \rightarrow z))$. The max function ensures that if at least one simple rule has a high conformity measure, then the complex rule has also a high conformity measure, i.e., $(K, \sqsubseteq)$ already contains the information carried out by R_2. Conversely, if all the simple rules have a low conformity measure, i.e., if there is no simple rule that is conform to the model $(K, \sqsubseteq)$, then R_2 is also considered as being not conform to $(K, \sqsubseteq)$.

n—m rules. Following the same two ideas, a n—m rule is considered as a conjunction of disjunction of simple rules. The 3—2 rule $R_3 : x \wedge y \wedge z \rightarrow v \wedge w$ can be decomposed into $[(x \rightarrow v) \vee (y \rightarrow v) \vee (z \rightarrow v)] \wedge [(x \rightarrow w) \vee (y \rightarrow w) \vee (z \rightarrow w)]$. Hence, the conformity for R_3 is defined by: $\min(\max(Cty(x \rightarrow v), Cty(y \rightarrow v), Cty(z \rightarrow v)), \max(Cty(x \rightarrow w), Cty(y \rightarrow w), Cty(z \rightarrow w)))$ and can be generalized for all simple and complex rules R into:

$$Cty(R : x_1 \wedge ... \wedge x_n \rightarrow y_1 \wedge ... \wedge y_m) = \min_{j=1}^{m}(\max_{i=1}^{n}(Cty(x_i, x_j)))$$

In doing so, we have to mention that the combination of min and max in the conformity measure for complex rules may lead to loose the fact that some keyterms for R, among all others, are related in $(K, \sqsubseteq)$.

Since other relations are absent in (K, ⊑), R should be presented to the analyst. This case is illustrated by the following rule r_{12}:

$$\begin{aligned} Cty(b \wedge c \rightarrow a \wedge e) &= \min(\max(Cty(b, a), Cty(c, a)), \max(Cty(b, e), Cty(c, e))) \\ &= \min((\max(0.03, 0.07), \max(\mathbf{0.37}, 0.07))) \\ &= \min(0.07, \mathbf{0.37}) = 0.07 \end{aligned}$$

3.5. A Small Example for Complex AR

Given (K, ⊑) in Figure 1, the Table 5 shows the classification of the 20 valid AR extracted – 16 complex and 4 simple – in an increasing order according to their conformity with (K, ⊑). We notice that the conformity classification for complex rules is, as we expected, different from the classification with the statistical measures given in Table 3. The difference is due to the use of an extra knowledge source (K, ⊑) for the former classification, rather than the text collection only as for the latter classification. The next section gives the main results of a qualitative analysis on real-word corpus. We follow the same principle as used for the simple example: by comparing conformity *versus* statistical measure classifications[1], and by considering the analyst's perspective on the appropriate knowledge units carried by the rules.

Table5.Conformity of the 20 AR in Table 2 with the model (K, ⊑) depicted in Figure 1

id	Rule	Conformity	id	Rule	Conformity
r_6	$b \rightarrow a \wedge c \wedge e$	0.03	r_{18}	$c \wedge e \rightarrow a \wedge b$	0.07
r_7	$e \rightarrow b \wedge c$	0.04	r_{20}	$c \rightarrow a$	0.07
r_{10}	$e \rightarrow a \wedge b \wedge c$	0.04	r_9	$a \rightarrow c \wedge d$	0.09
r_{19}	$e \rightarrow b$	0.04	r_4	$a \rightarrow b \wedge c \wedge e$	0.09
r_{13}	$d \rightarrow a \wedge c$	0.07	r_{11}	$a \rightarrow c$	0.09
r_{14}	$c \rightarrow b \wedge e$	0.07	r_2	$b \rightarrow c \wedge e$	0.19
r_{15}	$c \rightarrow a \wedge d$	0.07	r_3	$a \wedge b \rightarrow c \wedge e$	0.19
r_{16}	$c \rightarrow a \wedge b \wedge e$	0.07	r_8	$a \wedge e \rightarrow b \wedge c$	0.26
r_{17}	$c \wedge e \rightarrow b$	0.07	r_5	$b \wedge c \rightarrow e$	0.37
$\mathbf{r_{12}}$	$b \wedge c \rightarrow a \wedge e$	0.07	r_1	$b \rightarrow e$	0.37

4. APPLICATION ON MOLECULAR BIOLOGY CORPUS

4.1. Description of the Experiment

On the one hand, there is a corpus of 1361 scientific paper abstracts holding on molecular biology[2] of about 240,000 words (1.6 M-Bytes). The theme of the texts is the phenomenon of gene mutation causing a bacterial resistance to antibiotics. The interpretation results from this specific domain needs a high degree of human expertise. On the other hand, there is a domain ontology – a set of semantically related concepts – used as a knowledge model (K, ⊑). The concepts of the ontology are the correct keyterms of

[1] The statistical measure classification is detailed in [Cherfi *et al.*, 2006], where an algorithm is proposed and an evaluation is carried out by an analyst –expert in molecular biology.

[2] The corpus is collected from the Pascal-BioMed documentary database of the French institute for scientific and technical information (INIST)

the domain and constitute the pieces of information we mine in the texts. Moreover, we assume that cooccurrence of the keyterms in a text reflects a semantic link between keyterms [Anick & Pustejovsky, 1990]. We used UMLS [UMLS, 2000] restricted to the keyterms of the domain and all their parent keyterms represented by the specialisation relation (IsA). A keyterm is a noun phrase in the domain ontology which can be associated to a concept, and thus, it ensures the transition from the linguistic to the knowledge level. In this way, the corpus has been indexed with 14,374 keyterms, including 632 different keyterms. The minimal support σ_s for the AR extraction is set to 0.7% – occurring, at least in 10 texts – and the minimal confidence σ_c is set to 80%. We obtain 347 valid AR, including 128 exact rules. From the set of 347 rules, we kept 333 AR which do not deal with ambiguities in the keyterm meaning – two or more concept identifiers (CUI) in the UMLS for the same keyterm. Thus, we discarded 14 AR, and there are 510 different concepts remaining (from 632 original ones). When the 510 concepts are augmented with their IsA-parents, K is composed of 1,640 vertices (concepts) and 4178 edges ($\sqsubseteq$ relations). Among them, concepts appear 364 times in the 333 AR. There are 53 concepts in common with K (i.e., 56%), whereas 41 concepts are absent in K (i.e., 44%) out of the 94 different concepts from the AR set. There is a total number of 2,689,600 transitions probabilities computed from the 510 keyterms in K. The number of transition probabilities stored for the calculation in the 333 AR is: 419,906. The conformity computation operates 739 comparisons (min or max) for the probability transitions, yielding a total number of 831 values – with $108 \neq 0$ (i.e., 13%) and 21 different transitions, including $Cty = 0$. Finally, the conformity value range is [0, 0.231] with 18 different measure values and 75 out of 333 rules have their $Cty > 0$. We have to notice that the conformity measure is set to 0 for keyterms that does not appear in the $(K, \sqsubseteq)$ rather than a minimal probability as stated in section 3.2, because the automatic computation of the probability transitions for $(K, \sqsubseteq)$ is done once and regardless of the corpus. Finally, there are four classes of AR in the 333 set: 45 (1—1) simple rules (i.e., 13.51%), 5 (1—n) complex rules (i.e., 1.5%), 250 (n—1) complex rules (i.e., 75.08%), and 33 (n—m) complex rules (i.e., 9.9%). Table 6 summarizes these results.

Table6.Results on the model (K, $\sqsubseteq$) and the rule set extracted from our corpus

333 AR set	# concepts	# different concepts		
	364	94		
$(K, \sqsubseteq)$ model	# concepts	# concepts (Is-A augmented)		
	510	1640		
Transition probability	# values	# non-zero values		
	831	108 (13%)		
AR class	1—1	1—n	n—1	n—m
	45	5	250	33

4.2. Quality Analysis of AR rankings in the KBTM Process

The analysis is conducted as follows: For each rule in the four AR classes (1—n, 1—n, etc.), we compare its conformity measure, its statistical measures and whether or not it belongs to three groups based on the analyst's expertise: (i) interesting rules, (ii) relating known meronyms (especially hypernyms) and synonyms, and (iii) useless rules. Thanks to the conformity measure, we focus on a subset of 258 rules (*i.e.*, 77.5%) over the 333 rule set that are not conform to $(K, \sqsubseteq)$ – as they relate keyterms that either are absent in K or isolated concepts following the relation $\sqsubseteq$. This gives a significant improving rate of 22.5%

of extracted AR that are candidate to be discarded from the rule set. The discarded rules may be examined by their own in a further analysis (see summary in Table 7).

Table7.Results of the subset presented to the domain expert for qualitative analysis

AR category	# AR	Percentage (%)
interesting (Cty=0)	258	77.5
useless (Cty>0)	75	22.5
Total	333	

In the following, and without exhaustiveness, we report the outcome through some examples: Firstly, we focus on two close n–1 rules interesting according to the analyst: one is conform and the other is not conform with regards to (K, ⊑). Next, we show and comment one simple AR belonging to the analyst's class: relating known keyterms. We end with an example of a useless AR according to the analyst. Some rules are identified as interesting by the analyst. For example, the mutation of the parC gene is interesting to comment in the following two (2—1) rules:

Rule Number: 221
"gyra gene" ∧ "substitution" → "quinolone"
Interest: "13.610" Conviction: "4.706" Dependency: "0.741" Novelty: "0.008"
Satisfaction:"0.788" Conformity: "0"
Rule Number: 218
"gyra gene" ∧ "sparfloxacin" → "ciprofloxacin"
Interest: "1.073" Conviction: "6.003" Dependency: "0.770" Novelty: "0.007"
Satisfaction: "0.833" Conformity: "0.000215"

The rule #218, with Cty > 0, draws the resistance mechanism for two antibiotics sparfloxacin and ciprofloxacin that are subsumed (⊑) by the concept of quinolone (a family of antibiotics) in K. Moreover, the rule #221 is more precise by pointing out the specific resistance mechanism (namely substitution). We notice that the major good statistical measures for these rules are: conviction and satisfaction. Nevertheless, both measures give the reverse classification compared to the conformity and the analyst comments below. Some simple AR relate synonyms or hypernyms keyterms. They belong to the group: relating known keyterms according to the analyst. This group of rules shows that authors of the texts describe the same concept with different keyterms, and the text mining process reveals such usage.

Rule Number: 183:
"epidemic strain" −→ "outbreak"
Interest: "17.449" Conviction: "undefined" Dependency: "0.943" Novelty: "0.011"
Satisfaction: "1.000" Conformity: "0"

The statistical measure that gives a good quality for rule #183 is the dependency (which is used as the 3rd quality measure to check following the algorithm given in [Cherfi *et al.*, 2006]). The interest measure classes this rule in the middle of the corresponding list. Conversely, the conformity is 0, which gives it a chance to be analysed and update (K, ⊑) with two missing relations epidemic strain ⊑ outbreak and outbreak ⊑ epidemic strain.

Finally, the rules #268 and #269 are examples which are considered as wrong, hence useless for the analysis. It is due to the fact that keyterms: mycobacterium and tuberculosis are not significant in the molecular biology domain; however, these keyterms are extracted as keyterm index and are present as concepts in the general UMLS. The correct concept, in this context, would be the keyterm mycobacterium tuberculosis (see in [Cherfi *et al.*, 2006]).

Rule Number: 268
"mutation" $\wedge$ "mycobacterium tuberculosis" $\rightarrow$ "tuberculosis"
Interest: "14.956" Conviction: "undefined" Dependency: "0.933" Novelty: "0.006"
Satisfaction: "1.000" Conformity: "0.000178"
Rule Number: 269
"mutation" $\wedge$ "mycobacterium" $\rightarrow$ "tuberculosis"
Interest: "12.463" Conviction: "5.599" Dependency: "0.766" Novelty: "0.010"
Satisfaction: "0.821" Conformity: "0.00017809"

The rules #268 and #269 have the same non zero conformity, and have also good statistical quality measures. Hence, they will be presented to the analyst. Using the KBTM process, and without knowledge loss, we can discard the rules #268 and #269 from the rule set presented to the analyst because they are useless by introducing the artefacts mycobacterium and tuberculosis which are irrelevant in the context of molecular biology.

5. DISCUSSION

Among studies that intend to handle the large set of AR extracted with statistical quality measures, [Kuntz *et al.*, 2000] is similar to the work presented in section 2.2. This methodology is of great interest to highlight rule properties such as resistance to noise in the data set, or to establish whether a rule is extracted randomly or not (i.e., by chance). However, the limits of these measures come from the fact that they do not consider any knowledge model.

The background knowledge is used during the data mining process in [Jaroszewicz & Simovici, 2004] with a Bayesian Network [Pearl, 1988] to filter interesting frequent itemsets. A Bayesian network is similar to the knowledge model (K, $\sqsubseteq$) described in this chapter; except that each vertex (*i.e.*, relation) is associated with a weight defined by the relation conditional probability (*e.g.*, for the specialisation $\sqsubseteq$) *wrt.* to the concept parent(s) in the Bayesian network. The distribution probabilities over the relations are set up, *a priori*, by expert's judgments. The authors propose an algorithm to compute the marginal distributions of the itemset (*e.g.*, corresponding to the keyterm sets when dealing with text applications) over the Bayesian network. Hence, the itemset marginal distributions are inferred from the Bayesian network structure. An itemset is interesting if its support in the corpus (*i.e.*, real support of appearing in the texts) deviates, with a given threshold, from the support inferred from the Bayesian network (*i.e.*, its conditional probability to occur in the knowledge domain). A sampling-based approach algorithm for fast discovery of the interesting itemsets (called unexpected patterns) is given in [Jaroszewicz & Scheffer, 2005].

This methodology is extended in [Faure *et al.*, 2006] to drive both the AR extraction and the Bayesian network's weight updates. Hence, iteratively, the interesting AR identified in this way are candidates to update the Bayesian network. The similarities with the approach presented in this chapter are high.

However, when [Faure *et al.*, 2006] deal with probabilistic reasoning and analyst's judgments on the structure of the Bayesian Network, we rather stick to more formal knowledge conveyed by an ontological (i.e., consensual) domain knowledge model. However, the approach in [Faure *et al.*, 2006] could be complementary to the KBTM approach presented in this chapter. Further studies can be conducted to study the AR rankings given by both approaches for a given domain corpus *wrt.* to, respectively, a knowledge model, and a Bayesian network.

Another interesting work for the post-mining of association rules involving user interaction as backgroung knowledge is [Sahar, 1999; Liu *et al.,* 2003]. Here, the user is asked to interact with the system in order to evaluate the quality of the rules. [Sahar, 1999] assumes the following hypothesis: if a simple rule $k_1 \rightarrow k_2$ is of low interest for the user, then all related complex rules – related rules are defined as rules containing k_1 in their body and k_2 in their head– are also considered as of low interest. The user does not have to study them and the number of rules to study is substantially reduced. The user is asked to classify simple rules in one of the four categories: (1) true but uninteresting, (2) false and interesting, (3) false and uninteresting, (4) true and interesting. If a simple rule is classified in class (1) or (3), then the rule itself and its complex related rules may be deleted from the set of rules. This work has some other interesting characteristics: (i) An appropriate algorithm has been developed to select the simple rules to be given first to the user. The selected rules are the ones connected to a large number of complex rules. In this way, the number of rules to study decreases more rapidly than a random choice. (ii) The approach takes into account the direction of the rule: the deletion by the user of the rule $k_1 \rightarrow k_2$ has no effect on the rule $k_2 \rightarrow k_1$. (iii) [Sahar, 1999] does not use a knowledge model but the subjective judgement of the user which may be seen as an informal knowledge model. (iv) Finally, the major difference between our approach and [Sahar, 1999] concerns the interpretation of complex rules. The assumption adopted in [Sahar, 1999] is that any complex rule, according to our interpretation, could be turned to a conjunction of simple rules. However, we have shown that such decomposition, in clausal form, is misleading: $1 — m$ rules can be rewritten into a conjunction of simple rules; whereas $n — 1$ rules are rewritten into a disjunction of simple rules.

[Basu *et al.*, 2001] proposes another approach and uses WORDNET lexical network to evaluate the quality of the rule where keyterms are, actually, words. The quality score of a simple rule $word_1 \rightarrow word_2$ is given by the semantic distance between $word_1$ and $word_2$ in the lexical network. The network is a weighted graph, and each semantic relation (syno/antonymy, hyper/hyponymy) has its own weight. The distance between two words is the lower weight path in the graph. For any complex rule, the quality score is the mean of the distance for each pair $(word_i , word_j)$ where $word_i$ is in the body of the rule and $word_j$ is in its head. Here, as in [Sahar, 1999], the definition of the score for complex rules is logically false. The advantage in [Basu *et al.*, 2001] is the ability to deal with several semantic relations. However, the different properties of these relations cannot be formally expressed using a weighted graph and some assumptions are made such as: $weight(synonymy) > weight(hypernymy)$, etc. This method, based on a network of lexical entities, could be adapted to a formal knowledge model. However, it cannot be used to update a knowledge model: the weighting system and the mean calculation of the score for complex rules make impossible the association of a rule with a knowledge model as we did in Table 5.

6. CONCLUSION AND FUTURE WORK

In this chapter, we have proposed two methods for classifying association rules extracted within a KBTM process: the first one is based on statistical measures, and the second one is based on conformity with a knowledge model. Our present research study sets a knowledge-based text mining (KBTM) process driven by a knowledge model of the domain. Association rules that do not correspond to known relations of specialisation in the knowledge model are identified thanks to the conformity measure. The behaviour of the conformity measure is in agreement with the KBTM process. The conformity measure allows us both the enrichment of the knowledge model, and the TM process efficiency enhancement. An experiment on real-world textual corpus gives a significant improving rate and shows the benefits of the proposed approach to an analyst of the domain.

Furthermore, the conformity measure proposed in this first study can be extended to a number of promising directions in order to assess its effectiveness in different knowledge domains and contexts. Firstly, it could be interesting to take into account in the knowledge model of molecular biology domain other relations such as: causality (by considering rules involving instances of antibiotics → bacteria), temporal (the study of gene parC) mutation is anterior to gyrA study, how this relation has an impact on the resistance mechanism to antibiotics). In doing so, we will be able to have a deeper understanding of the texts and suggest an accurate modification of the knowledge model itself within the KBTM process.

REFERENCES

[Agrawal *et al.*, 1996] Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., & Verkamo, A. I. (1996). Fast Discovery of Association Rules. In U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, & R. Uthurusamy (Ed.), *Advances in Knowledge Discovery and Data Mining* (pp. 307–328). Menlo Park, CA: AAAI Press / MIT Press.

[Anick & Pustejovsky, 1990] Anick, P., & Pustejovsky, J. (1990). An Application of lexical Semantics to Knowledge Acquisition from Corpora. *30th International Conf. on Computational Linguistics (COLING'90)*: Vol. 3 (pp. 7–12). Helsinki, Finland.

[Bastide *et al.*, 2000] Bastide, Y., Taouil, R., Pasquier, N., Stumme, G., & Lakhal, L. (2000). Mining frequent patterns with counting inference. *ACM SIGKDD Exploration Journal*, 2(2): 66–75.

[Basu *et al.*, 2001] Basu, S., Mooney, R.J., Pasupuleti, K.V., and Ghosh J. (2001). Evaluating the Novelty of Text-Mined Rules using Lexical Knowledge. *7th ACM SIGKDD International Conference on Knowledge Discovery in Databases* (pp. 233–238). San Francisco, CA: ACM Press.

[Cherfi *et al.*, 2006] Cherfi, H., Napoli, A., & Toussaint, Y. (2006). Towards a text mining methodology using frequent itemsets and association rules. *Soft Computing Journal - A Fusion of Foundations, Methodologies and Applications,* 10(5):431–441. Special Issue on "Recent Advances in Knowledge and Discovery". Springer-Verlag.

[Collins & Loftus, 1975] Collins A. & Loftus E. (1975). A spreading-activation of semantic processing. *Psychological Review*, 82(6):407–428.

[Faure *et al.*, 2006] Faure, C., Delprat, D., Boulicaut, JF., & Mille, A. (2006). Iterative Bayesian Network Implementation by using Annotated Association Rules. *15th Int'l Conf. on Knowledge Engineering and Knowledge Management – Managing Knowledge in a World of Networks, Vol. 4248 of Lecture Notes in Artificial Intelligence – LNAI* (pp. 326–333). Prague, Czech Republic: Springer-Verlag.

[Fayyad *et al.*, 1996] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (1996). *Advances in Knowledge Discovery and Data Mining*. Menlo Park, CA: AAAI Press / MIT Press.

[Jacquemin, 2000] Jacquemin, C., (1994). FASTR: A unification-based front-end to automatic indexing. *Information multimedia, information retrieval systems and management* (pp.34–47). New-York, NY: Rockfeller University.

[Jaroszewicz & Scheffer, 2005] Jaroszewicz, S. & Scheffer, T. (2005). Fast Discovery of Unexpected Patterns in Data, Relative to a Bayesian Network. *ACM SIGKDD Conference on Knowledge Discovery in Databases* (pp. 118–127). Chicago, IL: ACM Press.

[Jaroszewicz & Simovici, 2004] Jaroszewicz, S. & Simovici, D.A. (2004) Interestingness of Frequent Itemsets using Bayesian networks as Background Knowledge. *ACM SIGKDD Conference on Knowledge Discovery in Databases* (pp. 178–186). Seattle, WA: ACM Press.

[Kuntz *et al.*, 2000] Kuntz, P., Guillet, F., Lehn, R., & Briand, H. (2000). A User-Driven Process for Mining Association Rules. D. Zighed, H. Komorowski, & J. Zytkow (Ed.). *4th Eur. Conf. on Principles of Data Mining and Knowledge Discovery (PKDD'00), Vol. 1910 of Lecture Notes in Computer Science – LNCS* (pp. 483–489), Lyon, France: Springer-Verlag.

[Lavrac *et al.*, 1999] Lavrac, N., Flach, P., & Zupan, B. (1999). Rule Evaluation Measures: A Unifying View. *9th Int'l Workshop on Inductive Logic Programming (ILP'99). Co-located with ICML'9., Vol. 1634 of Lecture Notes in Artificial Intelligence – LNAI (*pp. 174–185). Bled, Slovenia: Springer-Verlag, Heidelberg.

[Liu *et al.*, 2003] Liu, B., Ma, Y., Wong, C., & Yu, P. (2003). Scoring the Data Using Association Rules. *Applied Intelligence*, 18(2): 119–135.

[Pasquier *et al.*, 1999] Pasquier, N., Bastide, Y., Taouil, R., & Lakhal, L. (1999). Pruning closed itemset lattices for association rules. *International Journal of Information Systems*, 24(1): 25–46.

[Paroubek, 2007] Paroubek, P. (2007). Evaluating Part-Of-Speech Tagging and Parsing – On the Evaluation of Automatic Parsing of Natural Language (Chapter 4). In L. Dybkaer, H. Hemsen, & W. Minker (Ed.), *Chapter 4 of Evaluation of Text and Speech Systems* (pp. 99–124). Springer.

[Pearl, 1988] Pearl, J. (1988). Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. San Fransisco, CA: Morgan Kaufmann.

[Sahar, 1999] Sahar, S. (1999). Interestingness via What is Not Interesting. S. Chaudhuri, & D. Madigan, (Ed.), *5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'99)*. (pp. 332–336). San Diego, CA:. ACM Press.

[UMLS, 2000] UMLS (2000). *The Unified Medical Language System*., (11th edition): National Library of Medicine.

Annexe B

Fouille de données en pharmacovigilance par FCA [VTLL10]

Adverse Drug Reaction Mining in Pharmacovigilance data using Formal Concept Analysis

Jean Villerd[1], Yannick Toussaint[1], and Agnès Lillo-Le Louët[2]

[1] Loria – INRIA Nancy Grand Est, Nancy, France,
{firstname.lastname}@loria.fr,
[2] Pharmacovigilance Regional Center, Hôpital Européen G. Pompidou, Paris, France
agnes.lillo-lelouet@egp.aphp.fr

Abstract. In this paper we discuss the problem of extracting and evaluating associations between drugs and adverse effects in pharmacovigilance data. Approaches proposed by the medical informatics community for mining one drug - one effect pairs perform an exhaustive search strategy that precludes from mining high-order associations. Some specificities of pharmacovigilance data prevent from applying pattern mining approaches proposed by the data mining community for similar problems dealing with epidemiological studies. We argue that Formal Concept Analysis (FCA) and concept lattices constitute a suitable framework for both identifying relevant associations, and assisting experts in their evaluation task. Demographic attributes are handled so that the disproportionality of an association is computed w.r.t. the relevant population stratum to prevent confounding. We put the focus on the understandability of the results and provide evaluation facilities for experts. A real case study on a subset of the French spontaneous reporting system shows that the method identifies known adverse drug reactions and some unknown associations that has to be further investigated.

1 Introduction

Pharmacovigilance is the process of monitoring the safety of post-marketed drugs. The pharmacovigilance process starts with collecting *spontaneous case reports*: when suspecting an adverse drug reaction, health care practitioners send a case report to a spontaneous reporting system (SRS), mentioning the observed adverse effects, the drugs taken, and demographic data about the patient. These data are exploited by pharmacovigilance experts to detect *signals* of unexpected adverse drug reactions that require further clinical investigation. The size of these databases preclude their manual exploration: in 2008 more than 20,000 new cases were added to the French pharmacovigilance system while the WHO database contains more than 3 millions of reports.

The medical informatics community proposed some approaches that extract a set of *potential signals* for experts, i.e. a set of pairs (d, e) showing an unexpected correlation between an observed adverse effect e and the prescription

of a marketed drug d [1, 2]. Disproportionality measures have been introduced to quantify this notion of *unexpectedness* [3, 4]. However, the exhaustive search strategy performed by these approaches precludes from mining high-order associations between sets of drugs and adverse effects and from efficiently applying stratification on demographic attributes to prevent confounding.

In the meantime, the data mining community introduced statistical measures from epidemiology into the itemset and rule mining problems [5, 6]. Considering exposures as items and a given outcome as a class label, [7, 8] proposed efficient approaches that extract *risk patterns* (or *risk itemsets*) correlated with the given outcome. The relevance of a risk pattern is measured by statistical measures such as relative risk. Efficient pruning strategies have been proposed to reduce the search space and to provide concise representations of risk patterns. In particular, [9] considered optimal risk patterns where a risk pattern is said optimal if its relative risk is greater than the relative risk of all its subpatterns. This allows to reduce the number of extracted itemsets by discarding factors that do not increase the strength of shorter risk patterns.

However, some specificities of pharmacovigilance databases compared to epidemiological studies prevent from efficiently applying the above approaches. In contrary to epidemiological studies, pharmacovigilance databases are not designed to monitor one specific exposure to a drug or one specific outcome (adverse effect). Moreover, the database only contains situations "when things went wrong", leading to many potential biases that experts should take into account. In particular, demographic features may act as confounders and lead to extract spurious potential signals. Recent studies have shown that each demographic subpopulation should be separetely investigated by performing stratification [10]. This paper deals with the following issues that are currently not adressed by available tools from the medical informatics community :

1. **Dealing with demographic factors.** Stratification is not performed on demographic factors such as age and gender because exhaustively generating measures on all strata has a prohibitive cost. The aim at dealing with demographic factors is twofold. Firstly, it provides insights into the distribution by demographic factors for a given pair (d, e) and enables a comparative study. Secondly, demographic factors are used to guide further investigation such as clinical trials, especially in patients selection.
2. **Handling complex associations.** A signal of the form (d_1, e) can be related with more complex associations involving several drugs and several adverse effects. For example, if (d_1d_2, e) is recognised as a potential drug interactions, experts should be able to compare the respective strengths of (d_1d_2, e), (d_1, e), and (d_2, e).
3. **Providing a complete information.** Since pharmacovigilance data contain many sources of bias, a potential signal (d, e) should be presented to the experts only if there is no hidden additional factor shared by the corresponding group of patients that took d and suffered from e. For instance if this subgroup only contains men, (d, e, M), i.e. (d, e) on the male subpopulation, should be rather considered. Therefore, our aim is not to find the shortest

itemsets with the highest disproportionality, but to provide experts with potential associations (D, E, X) where the itemset DEX is the most complete description of the group of patients on which the potential association is observed.

In this paper, we propose a signal detection method based on Formal Concept Analysis that provides answers to these three points. Section 2 presents the issues concerning signal detection and introduces two constraints that define potential associations. Section 3 describes our method based on a concept lattice for identifying potential associations. Section 4 presents how the concept lattice provides features that help experts in evaluating potential interactions. An experiment on real data is analysed. Section 5 concludes the paper with a summary of contributions and future work.

2 Problem setting

Meyboom *et al.* [11] gives a comprehensive definition of signal detection process as being "A set of data constituting a hypothesis that is relevant to the rational and safe use of a medicine. Such data are usually clinical, pharmacological, pathological or epidemiological in nature. A signal consists of a hypothesis together with data and arguments." A *potential signal* is then an hypothesis suggested by an automated signal detection system that has to be evaluated by an expert. More precisely, a signal consists in (i) a pair (d, e) where d is suspected to be the cause of e (hypothesis), (ii) a set of reports (data), and (iii) disproportionality measures (arguments).

Only a few studies extended this definition to *potential associations*, i.e. higher-order hypothesis (D, E) where D and E are sets. have been published on higher-order associations, mainly about drug-drug interactions [12].

The aim of signal detection methods is to identify, among all pairs (d, e), those that occur more than expected when assuming the independance between d and e. However, although the number of reports for (d, e) is known in the database, the number of patients exposed to the drug d in the whole population is not, nor the number of patients suffering from e. Thus, the expected number of reports can not be reliably computed [13]. A solution consists in estimating the expected number of reports for (d, e) by considering the number of reports concerning other drugs and other adverse effects in the database. Therefore, contingency tables are central data structures. Table 1 depicts the contingency table for a pair (d, e). Each cell contains the number of reports corresponding to a given combination in the database: n_{11} is the number reports containing both d and e, i.e. the observed number of reports, n_{10} is the number of reports containing d but not e, and so on. N is the total number of reports. Several measures have been introduced to capture to what extent a pair is reported more than expected. The most widely used is the *Proportional Reporting Ratio (PRR)* [3], defined as

$$PRR(d, e) = \frac{P(e|d)}{P(e|\overline{d})} = \frac{\frac{n_{11}}{n_{11}+n_{10}}}{\frac{n_{01}}{n_{01}+n_{00}}}$$

4

The pair (d, e) is considered to be a potential signal if $PRR \geq 2$ and $\chi^2 \geq 4$ and $n_{11} \geq 3$ [3, 1]. This criterion is widely used, notably by the British Medicines and Healthcare products Regulatory Agency (MHRA). Intuitively, the first condition means that there must be twice as much probabilities to suffer from e while taking d, rather than while not taking d. The second one ensures that d and e are not independant. The third condition tells that there must be at least three reports containing d and e in the database. Other disproportionality measures such as the *Reporting Odds Ratio (ROR)* [4] are also used. More sophisticated methods implement disproportionality measures in a Bayesian framework [14].

Table 1. Contingency table for a signal (d, e).

	e	$\overline{e}$	
d	n_{11}	n_{10}	$n_{11} + n_{10}$
$\overline{d}$	n_{01}	n_{00}	$n_{01} + n_{00}$
	$n_{11} + n_{01}$	$n_{10} + n_{00}$	N

Table 2. Contingency table on a subpopulation.

	eM	$\overline{e}M$	
dM	n_{11}	n_{10}	$n_{11} + n_{10}$
$\overline{d}M$	n_{01}	n_{00}	$n_{01} + n_{00}$
	$n_{11} + n_{01}$	$n_{10} + n_{00}$	$supp(M)$

Demographic factors such as gender and age may help in identifying vulnerable subpopulations. Indeed, drugs may be administered differentially according to age (e.g. vaccines), gender, or both of them (e.g. contraceptive pills), and some adverse effects may only concern a specific subpopulation (e.g. sudden infant death syndrome). Therefore, disproportionality should be computed on groups of patients that belong to the same subpopulation. This stratification process leads to compute a PRR_{strat} value on each subpopulation, called *stratum*, for a given pair (d, e). For instance, the PRR_{strat} of (d, e) on the male subpopulation is $PRR_{strat}(d, e, M) = \frac{P(e|dM)}{P(e|\overline{d}M)}$ computed from a contingency table where each cell is restricted to the male subpopulation (see Table 2 where $supp(M)$ is the number of male patients). Similarly, $\chi^2_{strat}(d, e, M)$ denotes the χ^2 value computed from the restricted contingency table. Experts compare PRR_{strat} values between strata to evaluate if the strength of the association between d and e depends on a demographic factor. For instance, if (d, e) has the same PRR_{strat} value on both male and female strata, gender is not an increasing factor.

Stratification also allows to detect situations where demographic factors act as confounders [10]. Unbalanced subpopulations may lead to situations where $PRR_{strat}(d, e, M)$ and $PRR_{strat}(d, e, F)$ are equals while $PRR_{strat}(d, e, \emptyset)$ (w.r.t. the whole population) has a different value. In such case, $PRR_{strat}(d, e, \emptyset)$ is not reliable and is said to be counfounded by gender. Therefore both *crude* $PRR_{strat}(d, e, \emptyset)$ and $PRR_{strat}(d, e, x_i)$ on strata x_i are relevant for experts to evaluate the strength and the reliability of a signal (d, e).

The three initial issues mentioned in introduction can be refined in extracting potential associations (D, E, X) such that:

1. the disproportionality of (D, E, X) is computed w.r.t. the subpopulation X, following the stratification strategy;

2. potential associations are presented to the experts in such a way that comparisons between an association (d_1, e, M) and its related associations (e.g. (d_1d_2, e, M), $(d_1, e, \emptyset)$) is straightforward;
3. considering a potential association (D, E, X), the corresponding group of patients do not share any additional attribute than those in DEX.

3 A FCA-based signal detection method

Let $\mathcal{D}$ be a set of drugs, $\mathcal{E}$ be a set of adverse effects and $\mathcal{X}$ a set of binarized demographic attributes. We look for potential associations (D, E, X), ($D \subseteq \mathcal{D}$, $E \subseteq \mathcal{E}$, $X \subseteq \mathcal{X}$, $D \neq \emptyset$, $E \neq \emptyset$) that satisfy two types of constraints:

- a closure constraint: stating that patients that cover the itemset $D \cup E \cup X$, noted DEX, do not share any additional attribute,
- a strength constraint: stating that $supp(DEX) \geq 3$, $PRR_{strat}(D, E, X) \geq 2$ and $\chi^2_{strat}(D, E, X) \geq 4$.

The closure constraint clearly says that DEX must be a closed itemset. Thus, our search space for potential associations consists of closed itemsets that contain at least one element of $\mathcal{D}$ and one element of $\mathcal{E}$. In the following we present basics on Formal Concept Analysis and concept lattices. We later show that the concept lattice is a suitable structure for extracting potential associations, in the sense that it covers our search space, and that it provides experts with efficient ways of comparing related associations.

3.1 Basics on Formal concept analysis

Considering a binary relation between a set of objects $\mathcal{O}$ and a set of binary attributes $\mathcal{A}$, FCA extracts a set of pairs (O, A) with $O \subseteq \mathcal{O}$, $A \subseteq \mathcal{A}$, called formal concepts, such that each object in O owns all attributes in A and vice-versa. Formal concepts are partially ordered w.r.t. the inclusion of O and A, to form a lattice structure called concept lattice. In that way, the concept lattice can be seen as a conceptualization of the binary relation.

In the following, we present formal definitions from [15]. A *formal context* is a triple $\mathbb{K} = (\mathcal{O}, \mathcal{A}, I)$ where $\mathcal{O}$ is a set of *objects*, $\mathcal{A}$ a set of *attributes*, and $I \subseteq \mathcal{O} \times \mathcal{A}$ a binary relation such that oIa if the object o owns the attribute a. Figure 1 shows a formal context $\mathbb{K}$ with $\mathcal{O} = \{o_1 \dots o_7\}$ and $\mathcal{A} = \{d_1 \dots d_3\} \cup \{e_1, e_2\} \cup \{M, F\}$.

Two *derivation operators*, both denoted by $(.)'$, link objects and attributes. Considering a set of objects $O \subseteq \mathcal{O}$, $O' = \{a \in \mathcal{A} | oIa\}$, i.e. O' is the set of attributes shared by all objects in O. Dually, $A' = \{o \in \mathcal{O} | oIa\}$ is the set of objects that own all attributes in A. $|A'|$ is called the *support* of A, noted $\sigma(A)$. For instance, $\{d_1, d_2\}' = \{o_3, o_4\}$ and $\{o_3, o_4\}' = \{d_1, d_2, e_1, M\}$.

Two compound operators, both denoted by $(.)''$, composed of the two previous derivation operators, are *closure operators* on $2^{\mathcal{O}}$ and $2^{\mathcal{A}}$. Therefore O'' is the maximal set of objects that share the same attributes than the objects in O.

Dually, A'' is the maximal set of attributes that are owned by the objects that share attributes in A. A set of attribute A is said to be *closed* if $A = A''$. The set of sets B such that $B'' = A$ forms the *equivalence class* of A. All sets in the equivalence class of A have the same support $\sigma(A)$. For instance, $\{d_1, d_2\}$ is not closed since $\{d_1, d_2\}'' = \{o_3, o_4\}' = \{d_1, d_2, e_1, M\}$, while $\{o_3, o_4\}$ is closed since $\{o_3, o_4\}'' = \{d_1, d_2, e_1, M\}' = \{o_3, o_4\}$.

A *formal concept* is a pair (O, A) such that $O = O''$ and $A = O'$. Each object in O owns all attributes in A and vice-versa. Both O and A are closed sets, which means that no object (resp. attribute) can be added to O (resp. A) without changing A (resp. O). O (resp. A) is called the *extent* noted $\text{Ext}(O, A)$ (resp. the *intent* noted $\text{Int}(O, A)$) of the concept. The set of all formal concepts of the formal context $\mathbb{K}$ is denoted $\mathfrak{B}(\mathbb{K})$. For instance, $(\{o_3, o_4\}, \{d_1, d_2, e_1, M\})$ is a formal concept.

Formal concepts are partially ordered w.r.t. to the inclusion of their extents. Considering two concepts (O_1, A_1) and (O_2, A_2), $(O_1, A_1) \leq (O_2, A_2)$ iff $O_1 \subseteq O_2$ (which is equivalent to $A_1 \supseteq A_2$). The set of all formal concepts ordered in this way is denoted by $\underline{\mathfrak{B}}(\mathbb{K})$ and is called the *concept lattice* of the formal context $\mathbb{K}$. The maximal concept $(\mathcal{O}, \mathcal{O}')$ is called the *top* concept, and the minimal concept $(\mathcal{A}', \mathcal{A})$ is called the *bottom* concept.

The concept lattice $\underline{\mathfrak{B}}(\mathbb{K})$, built from $\mathbb{K}$ is shown in Figure 1. Each box represents a formal concept with its intent in the upper part, and its extent in its lower part.

Considering an attribute a, its *attribute concept*, denoted $\mu(a)$, is the unique concept (a'', a'), i.e. the highest concept that contains a in its intent on Figure 1. For instance, $\mu(e_2) = (\{o_1, o_7\}, \{e_2\})$.

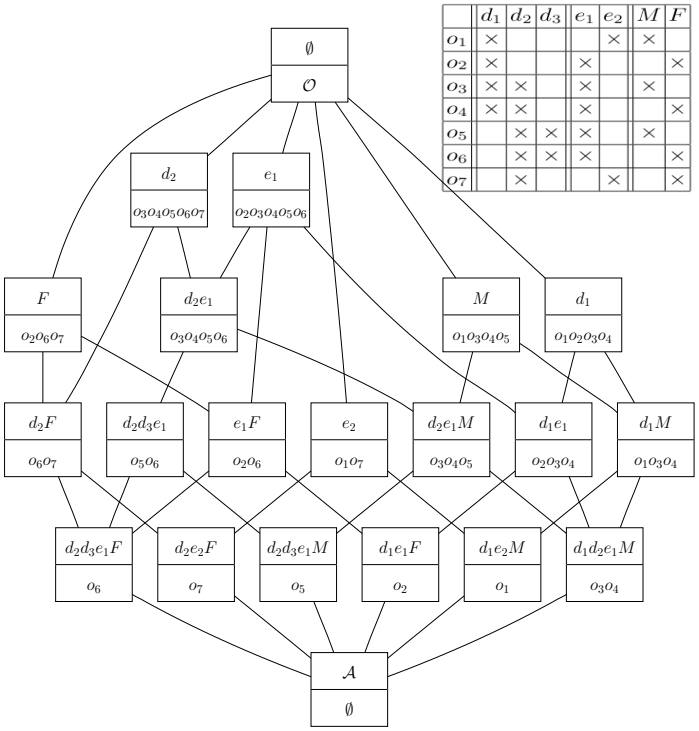
In the worst case, the number of concepts of $\mathbb{K} = (\mathcal{O}, \mathcal{A}, I)$ is $2^{\min(|\mathcal{O}|, |\mathcal{A}|)}$. This occurs when each subset of $\mathcal{O}$ or $\mathcal{A}$ is closed, which is improbable in practice.

3.2 Our approach

Our aim is to extract potential associations that satisfy a closure constraint and a strength constraint. We showed that only closed itemsets can satisfy these constraints. Moreover our aim is to provide an understandable representation of results. As said before, interpretation is a difficult task for experts since pharmacovigilance data may contain many biases. Since the content of the database is not the result of a sampling method, spurious potential associations may be extracted. Disproportionality measures can not make the difference between a spurious disproportion due to a selection bias and a real disproportion due to an adverse effect reaction. Only experts can make this difference w.r.t. the content of the database and their domain knowledge. Therefore, in order to evaluate a potential association (D, E, X), experts need more information than disproportionality measures. They need to put back the association in its context of extraction, i.e. in the portion of the database where the disproportionality occurs.

The concept lattice is then a suitable structure for signal detection. It is built from the context $(\mathcal{O}, \mathcal{A}, I)$, where $\mathcal{O}$ is the set of reports, and $\mathcal{A} = \mathcal{D} \cup \mathcal{E} \cup \mathcal{X}$

7



	d_1	d_2	d_3	e_1	e_2	M	F
o_1	×				×	×	
o_2	×			×			×
o_3	×	×		×		×	
o_4	×	×		×			×
o_5		×	×	×		×	
o_6		×	×	×			×
o_7		×			×		×

Fig. 1. A formal context and its associated concept lattice

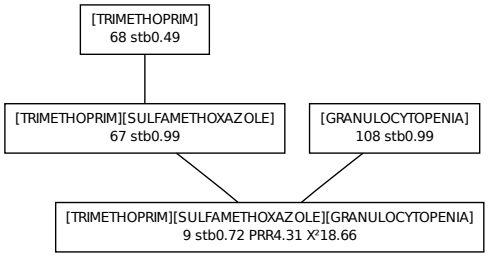


Fig. 2. An interaction example containing noise

8

is the set of attributes. Since concept intents are closed itemsets, the search space for potential associations is the set of concepts. Moreover, considering an association (d_1d_2, e_1, M), the partial order between concepts allows to isolate relevant information for interpretation that will be presented to experts: more specific strata among descendants of the concept with intent $\{d_1, d_2, e, M\}$, more general strata among ascendants for instance. But also to compare strengths of related associations: more specific associations (e.g. (d_1d_2, e_1e_2, M)) will be found among descendants and more general among ascendants.

Thus, our algorithm for extracting potential associations is straightforward. Concepts whose intent contains at least one drug and one adverse effect, and whose extent contains at least three reports are considered as candidate associations. Their contingency table is computed w.r.t. the demographic attributes in intent. If the MHRA criterion is satisfied, the intent is added to the set of potential associations.

Data: a concept lattice $\mathcal{L}$
Result: a set of potential associations P
foreach *concept* $c \in \mathcal{L}$ **do**
  if *Int(c) contains at least one element of* $\mathcal{D}$ *and one element of* $\mathcal{E}$ *and* $|Ext(c)| \geq 3$ **then**
    compute the contingency table for Int(c)
    compute PRR_{strat} and χ^2_{strat} values from the contingency table
    if $PRR_{strat} \geq 2$ *and* $\chi^2_{strat} \geq 4$ **then**
      add Int(c) to the set of potential associations P
    end
  end
end

Algorithm for signal detection

Therefore, the number of candidate associations is bounded by $2^{\min|\mathcal{O}|,|\mathcal{A}|}$, which is the number of formal concepts in the worst case. In practice, the number of reports is larger than the number of attributes, and all subsets of $\mathcal{A}$ are not closed.

Computing contingency tables Contingency tables are built from the lattice, in order to compute PRR_{strat} and χ^2_{strat} values.

Since each candidate association (D, E, X) is a closed itemset, there exists a unique formal concept c_{DEX} with $\text{Int}(c_{DEX}) = D \cup E \cup X$. We show in the following that the contingency table of any association can be computed knowing the support of c_{DEX} and the extent of the attribute-concepts $\mu(a), a \in DEX$.

In the general case of an association (D, E, X), The cell values of its contingency table restricted to the subpopulation X are computed as follows.

$$n_{11} = \sigma(DEX) = |\bigcap_{a \in DEX} \text{Ext}(\mu(a))| = |\text{Ext}(c_{DEX})|$$
$$n_{10} = \sigma(D\overline{E}X) = \sigma(DX) - \sigma(DEX) = |\bigcap_{a \in DX} \text{Ext}(\mu(a))| - n_{11}$$
$$n_{01} = \sigma(\overline{D}EX) = \sigma(EX) - \sigma(DEX) = |\bigcap_{a \in EX} \text{Ext}(\mu(a))| - n_{11}$$
$$n_{00} = \sigma(\overline{D}\,\overline{E}X) = |\bigcap_{a \in X} \text{Ext}(\mu(a))| - (n_{11} + n_{10} + n_{01})$$

Insights for noise detection The concept lattice provides an additional measure that helps in evaluating the reliability of a potential association. The *stability index* of a formal concept [16] quantifies the ability of the concept to remain existent after deletion of objects in its extent. In other words, the stability index of a concept c is low if $\text{Int}(c)$ becomes non-closed after the removal of a few objects from $\text{Ext}(c)$. Then, an unstable concept c correspond to a *barely closed* itemset $\text{Int}(c)$. Therefore, stability can be presented to experts as an additional quality measure for potential association. A potential association (D, E, X) with a low stability index barely satisfies the closure constraint and should be considered with care by experts.

Moreover, stability can provide insights for detecting noisy reports. We illustrate this aspect on a real example. Trimethoprim (d_1) and sulfamethoxazole (d_2) come together in the dosage form of marketed drugs, thus a unique concept $\mu(d_1) = \mu(d_2)$ should exist in the lattice. It is not the case (see Figure 2) since $\mu(d_2) \leq \mu(d_1)$ and $\sigma(\mu(d_1)) = 68$ while $\sigma(\mu(d_2)) = 67$. This means that, among all patients that took d_1, only one did not take d_2, which probably correspond to a badly filled report. The stability index can capture such a situation. Here, $\mu(d_1)$ has a low stability since the removal of the noisy report will lead d_1 to become non-closed with $d_1'' = \{d_1, d_2\}$ and then $\mu(d_1)$ will become $\mu(d_1) = \mu(d_2)$. Thus, a low stability index for a given concept should draw experts' attention to the potentially noisy reports contained in its extent.

3.3 Related works

Several works focused on finding *risk patterns* in epidemiological studies. Considering a set of patients described by a set of nominal attributes, and a target outcome e that partition patients into two classes (presence/absence), a risk pattern is a set of attribute-value pairs D such that the pattern is locally frequent ($support(De) \geq min\_sup$) and its *relative risk* is higher than a given threshold. Relative risk $RR(D, e) = \frac{P(e|D)}{P(e|\overline{D})}$ is a widely used measure in epidemiological studies. Note that PRR and RR formula are identical when ignoring demographic factors. [9] proposed algorithms for efficiently mining risk patterns. A

risk pattern is said *optimal* if its relative risk is greater than the relative risk of all its subpatterns. This allows to reduce the number of extracted patterns by discarding factors that do not increase the strength of more general risk patterns.

Although PRR and RR formula are identical for a given outcome e and a set of attributes D, this approach do not fit well our requirement for pharmacovigilance.

First there is no predefined outcome in pharmacovigilance data. Each combination of adverse effects may be considered as an outcome. Applying the precited approach would consist in generating the set of optimal risk patterns for each combination of adverse effects. This also prevents from applying other approaches such as subgroup discovery [17] and contrast set mining [18].

Secondly, in [9] demographic attributes play the same role as drugs in contigency tables. This means that the PRR of the pattern $\{d, M\}$ is computed as $PRR(d, e, M) = \frac{P(e|d,M)}{P(e|\overline{d},M)}$. In order to be consistent with the stratification recommendation about demographic factors, the PRR of $\{d, M\}$ should be computed w.r.t. the male stratum. It should compare men that took d and suffered from e with men that did not take d and suffered from e: $PRR_{strat}(d, e, M) = \frac{P(e|d,M)}{P(e|\overline{d},M)}$.

Thirdly, as defined in [9], risk patterns may not be closed itemsets and therefore may not satisfy our closure constraint.

Suppose that d_1d_2e is a closed itemset and that the closure of d_1 is $d_1'' = d_1d_2$, then $PRR(d_1, e, \emptyset) = PRR(d_1d_2, e, \emptyset)$ as well as $PRR_{strat}(d_1, e, \emptyset) = PRR_{strat}(d_1d_2, e, \emptyset)$. The risk pattern d_1d_2 is not optimal since its PRR value is not higher than its subpattern d_1 and is not retrieved, while following our constraints d_1d_2 has to be retrieved and not d_1. Moreover, the fact that risk patterns are extracted w.r.t. a given outcome would lead to generate risk patterns for e_1 and then risk patterns for e_2 without paying attention to situations where $e_1'' = e_2$. In this case, risk patterns w.r.t. e_1 do not satisfy our closure constraint since all patients that suffer from e_1 also suffer from e_2.

Moreover, considering non-closed itemsets prevents from computing PRR_{strat} in an accurate way. Consider the group of patients that took a drug d. Suppose that all patients that took d are men, i.e. the closure of $\{d\}$ is $\{d, M\}$. Then $PRR_{strat}(d, e, \emptyset) = \frac{P(e|d,\emptyset)}{P(e|\overline{d},\emptyset)} = \frac{P(e|d,M)}{P(e|\overline{d},\emptyset)}$. The numerator group of patients actually belongs to a more specific stratum (men) than the denominator group (men and women). $PRR_{strat}(d, e, \emptyset)$ can not be reliably computed w.r.t. the available data since only men took d. In this case, associations involving d are only reliable w.r.t. the male subpopulation. No reliable hypothesis can be made about d and e on the whole population since there is no female subpopulation that would allow to evaluate if (d, e) depends on gender or not.

Since signal detection aims at providing experts with hypothesis for further investigation, we claim that the reliability of an hypothesis is at least as important as its statistical strength. An hypothesis (D, E, X) is reliable if the corresponding set of patients do not share an additional attribute that may delude experts, i.e. if DEX is a closed itemset. This is true for demographic attribute as shown before but also for drugs and adverse effects.

4 Evaluation facilities and experimentation

This section shows how experts get a *contextualized* association using our approach. In addition to disproportionality measures, insights are given to help them in deciding whether a signal or an interaction should be further investigated or not.

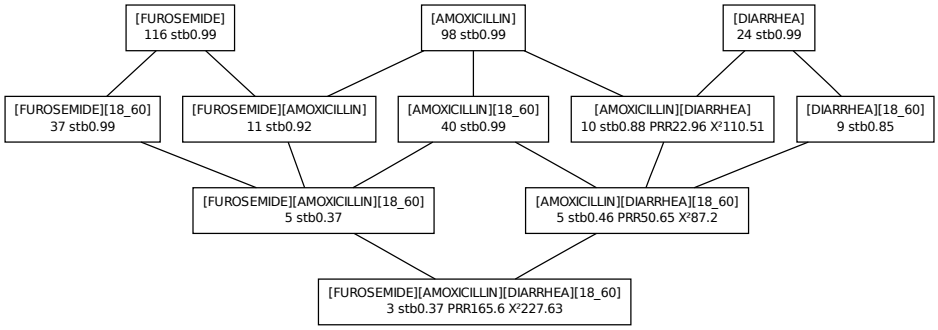


Fig. 3. Subpart of the lattice illustrating a potential interaction

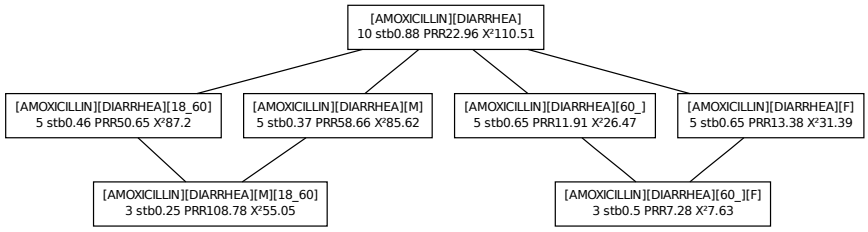


Fig. 4. Comparison of the different strata of a potential signal

4.1 Visualization and navigation

The core idea is to use the concept lattice as a synthetic representation of the database. From the list of potential association, experts access to a detailed

view that shows a subpart of the concept lattice, revealing additional information compared to statistical measures and helping experts in their interpretation and evaluation task.

Figure 3 shows the user interface illustrating a potential interaction (d_1d_2, e, X) where d_1 is `amoxicillin`, d_2 is `furosemide`, e is `diarrhea` and X is `18_60`, meaning age between 18 and 60.

A subpart of the lattice is shown, which contains the concept $c_{d_1d_2eX}$, corresponding to the interaction, at the bottom, the attribute-concepts $\mu(d_1)$, $\mu(d_2)$, $\mu(e)$ at the top, and all concepts on the paths from $c_{d_1d_2eX}$ to the attribute-concepts. Then the graph shows concepts that are more general than $c_{d_1d_2eX}$.

Concepts are labeled with their intent, support and stability. Concepts that own at least one drug and one adverse effect are also labelled with PRR_{strat} and χ^2_{strat} values. Through this graph, experts can compare the PRR_{strat} values of the interaction (d_1d_2, e, X) with those of the signals (d_1, e, X) and $(d_1, e, \emptyset)$, and observe that there are no concepts representing the signals (d_2, e, X) and $(d_2, e, \emptyset)$. This gives the information that no patient took `furosemide` and suffered from `diarrhea` without `amoxicillin`. Concepts that do not correspond to associations are also relevant. For instance, experts can observe that among the 24 patients that suffered from `diarrhea`, 10 took `amoxicillin` and state whether this ratio is realistic or is due to a selection bias.

Another graph (cf. Figure 4) shows a given association as root and those of its subconcepts that correspond to its demographic strata with a least 3 reports. Experts can compare their respective PRR_{strat} values and observe that, in this example, age distribution is different in male and female strata.

4.2 Experimentation

We applied our method on a subset of the French national SRS database. This subset contains 3249 cases, 976 drugs, 573 adverse effects. Two demographic attributes, gender and age are binarized into 6 binary attributes (2 for gender and 4 for age). The resulting lattice contains 13178 concepts, among which 6788 with support ≥ 3. Since only signals (one drug, one adverse effects), and interactions (two drugs, one adverse effect) are currently considered by pharmacovigilance experts, we only showed potential signals and interactions to experts. The 2812 candidate signals led to 786 potential signals and the 836 candidate interactions to 183 potential interactions.

Review of potential signals Potential signals were reviewed by an experts who classified them into 5 categories (see Table 3). Categories (1),(2) contain true positives, (3),(4) false positives and (5) unknown potential signals. 27 signals were classified as unknown, i.e. not reported in the literature, but interesting enough for further investigation by experts.

True positives are consistent with results of previous studies [19] and no known true-positive is missing. In the majority of cases, the demographics attributes associated to the couple drug/effect constitute a known risk factor or

probable risk factor. For example, cases of `Pulmonary Hypertension` associated with the use of appetite suppressants amphetamine-like were observed in women, between the ages of 18 and 60.

False positives (contained in categories (3) and (4)) are common in signal detection and some of them are well-known. The signal (`hydrochlorothiazide`, `cough`) is detected because these drug and adverse effects often appear together. However in these cases, cough is actually caused by ACE inhibitors taken concomitantly with `hydrochlorothiazide`. Since there are several ACE inhibitors d_i, each association (d_i,`cough`) appears with a lower support than the association (`hydrochlorothiazide`, `cough`), which may delude experts. A solution would be to introduce drug therapeutic families, such as ACE, as attributes, with (o, `ACE`) $\in I$ for each case o containing an ACE inhibitor. Then signals of the form (`ACE`, `cough`) would be detected, where ACE is a drug family, even if each signal (d, e) where d is an ACE inhibitor is too rare to be detected. Current improvements of our method aim at solving this problem.

Table 3. Potential signals

category	count	
1. known (in reference documents)	720 (91.6%)	true positives
2. known (in a similar form)	24 (3.1%)	
3. the effect is the origin of the medication	3 (0.4%)	false positives
4. due to concomitant drug	11 (1.4%)	
5. unknown potential signal	28 (3.5%)	further investigations needed

Review of potential interactions The evaluation of interactions is more difficult since it involves complex pharmacokinetics aspects. Moreover there is no consensus on whether (d_1d_2, e, X) should be considered as an interaction when both d_1 and d_2 are known to be the cause of e. Thus, we are not able to separate true and false positives. Experts classified the 183 potential interactions into 4 categories (see Table 4). The last category correspond to cases where further investigations are needed.

Table 4. Potential interactions

category	count
either d_1 or d_2 is a known cause of e	64(35.0%)
both d_1 or d_2 are known causes of e	66(36.0%)
d_1 and d_2 in the same dosage form	34(18.6%)
neither d_1 or d_2 are known causes	19(10.4%)

We noted that, in some cases, the PRR_{strat} value of an interaction (d_1d_2, e, X) where only d_1 is a known cause of e was greater than $PRR_{strat}(d_1, e, X)$. In such

cases, it is not clear if the focus should be put on (d_1d_2, e, X) or on (d_1, e, X). To our knowledge, there have been no pharmacovigilance study on defining preferences between an interaction (d_1d_2, e, X) and a signal (d_1, e, X) w.r.t. PRR value. Therefore, we can not discard (d_1d_2, e, X) when $PRR_{strat}(d_1d_2, e, X) < PRR_{strat}(d_1, e, X)$. This prevents from using the pruning strategy of the optimal risk patterns approach [9], that would discard (d_1d_2, e, X).

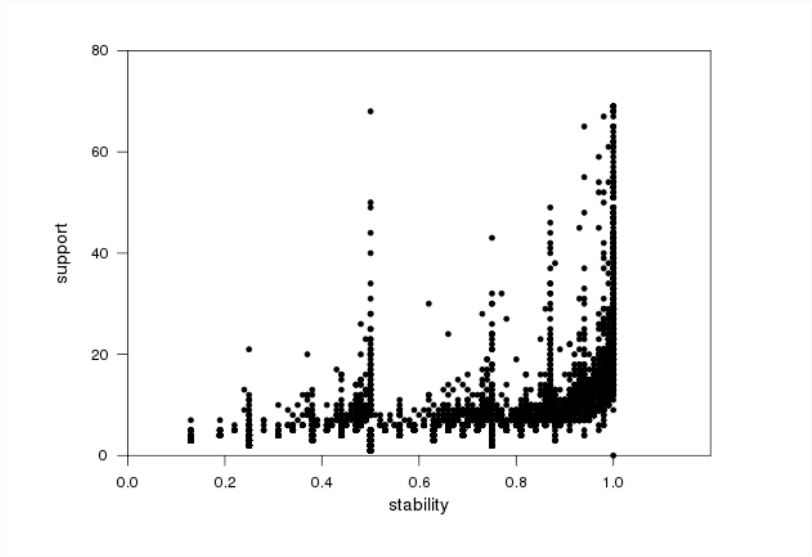


Fig. 5. Stability and support

Detection of noisy reports In a previous section, we showed that the stability index of a concept may be a clue for noisy reports detection. However, we faced the difficulty of defining a threshold on stability that defines *unstable* concepts. Frequent unstable concepts are interesting. They can be seen as concepts that gather a high number reports, but that actually exist because of only a few of them, which may be noisy reports. Frequent unstable concepts should be found in the upper left hand corner of the Figure 5. We empirically decided to investigate the 20 concepts with a minimum support of 20 reports and a stability index below 0.5. All of these concepts were in the same configuration than in Figure 2, i.e. among the n reports gathered by the unstable concept, $n - 1$ also share another attribute. For instance, among the 20 reports gathered by the unstable concept with intent {`tacrine`, `M`}, 19 also own the attribute $age > 60$. Since tacrine is used in the treatment of Alzheimer's disease, the report that does not own $age > 60$ is suspect and should be verified. The expert considered that the n^{th} report was actually suspect in 19 of the 20 unstable concepts under review.

5 Conclusion

In this paper, we presented an automated signal detection method, based on concept lattices, that provides a framework for extracting potential associations and performing qualitative analysis of the extracted associations. Potential associations are identified w.r.t. the MHRA criterion.

We claim that only associations that are closed itemsets should be presented to experts, since non-closed associations do not fully describe the set of factors shared by a subgroup of patients. Demographic attributes are taken into account in the PRR computation so that the disproportionality of an association is computed w.r.t. the subpopulation in which the association is observed. The closure constraint allows to identify the accurate subpopulations and prevents from exhaustively evaluate each population stratum.

Our method is thought for extracting complex associations, i.e. extracting associations where there are one or more drugs, one or more adverse effects and several demographic factors. Nowadays, if signals have been quite well studied, little work has been done on interactions, and practically none on syndromes (1 drug, several effects) or protocoles (several drugs, several effects) which justifies the facts that our evaluation has only been performed on signal and interactions.

When evaluating extracted associations, experts have access to subparts of the lattice for visualizing related associations, for example, an interaction is displayed with its related signals as well as its different "strength" on subpopulations. This visualization is of particular interest when both a signal or an interaction pass the MHRA criterion. Only experts – no automated process – are able to decide which of signals and interactions should be validated, mostly because of pharmacokinetics complexity. The interface is designed to facilitate a qualitative analysis by experts and guides exploration, interpretation and validation of associations.

Evaluation has been succesfully performed on the HEGP database. All already known signals have been found, and 28 signals need further investigations in literature and clinical trials as well as 19 unknown interactions. Evaluation was peformed by two experts separately who enjoyed visualization facilities.

Finally, we are currently improving our method for dealing with new issues, especially, for dealing with families of drugs and adverse effects.

References

1. Hauben, M., Madigan, D., Gerrits, C.M., Walsh, L., Puijenbroek, E.P.V.: The role of data mining in pharmacovigilance. Expert Opinion on Drug Safety **4**(5) (September 2005) 929–948
2. Bate, A., Lindquist, M., Edwards, I.R.: The application of knwoledge discovery in databases to post-marketing drug safety: example of the WHO database. Fundamental & Clinical Pharmacology **22**(2) (April 2008) 127–140
3. Evans, S.J.W., Waller, P.C., Davis, S.: Use of proportional reporting ratios for signal generation from spontaneous adverse drug reaction reports. Pharmacoepidemiology and Drug Safety **10**(6) (October/November 2001) 483–486

16

4. van der Heijden, P.G.M., van Puijenbroek, E.P., van Buuren, S., van der Hofstede, J.W.: On the assessment of adverse drug reactions from spontaneous reporting systems: the influence of under-reporting on odds ratios. Statistics in Medicine **21**(14) (July 2002) 2027–2044
5. Morishita, S., Sese, J.: Transversing itemset lattices with statistical metric pruning. In: Proceedings of the nineteenth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, ACM (2000) 226–236
6. Gu, L., Li, J., He, H., Williams, G., Hawkins, S., Kelman, C.: Association rule discovery with unbalanced class distributions. AI 2003: Advances in Artificial Intelligence 221–232
7. Li, H., Li, J., Wong, L., Feng, M., Tan, Y.: Relative risk and odds ratio: A data mining perspective. In: Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, ACM (2005) 377
8. Li, J., Fu, A., He, H., Chen, J., Jin, H., McAullay, D., Williams, G., Sparks, R., Kelman, C.: Mining risk patterns in medical data. In: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, ACM (2005) 770–775
9. Li, J., Fu, A., Fahey, P.: Efficient discovery of risk patterns in medical data. Arificial Intelligence in Medicine **45**(1) (2009) 77–89
10. Woo, E., Ball, R., Burwen, D., Braun, M.: Effects of stratification on data mining in the US Vaccine Adverse Event Reporting System (VAERS). Drug safety **31**(8) (2008) 667–674
11. Meyboom, R.H., Egberts, A.C., Edwards, I.R., Hekster, Y.A., De Koning, F.H.P., Gribnau, F.W.J.: Principles of signal detection in pharmacovigilance. Drug Safety **16**(6) (June 1997) 335–365
12. Almenoff, J., DuMouchel, W., Kindman, L., Yang, X., Fram, D.: Disproportionality analysis using empirical Bayes data mining: a tool for the evaluation of drug interactions in the post-marketing setting. Pharmacoepidemiology and Drug Safety **12**(6) (2003) 517–521
13. Roux, E., Thiessard, F., Fourrier, A., Bégaud, B., Tubert-Bitter, P.: Evaluation of statistical association measures for the automatic signal detection generation in pharmacovigilance. IEEE Transactions on Information Technology in Biomedicine **9**(4) (December 2005) 518–527
14. DuMouchel, W.: Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. The American Statistician **53**(3) (August 1999) 177–190
15. Ganter, B., Wille, R.: Formal concept analysis: Mathematical Foundations. Springer, Berlin (1999)
16. Kuznetsov, S.: On stability of a formal concept. Annals of Mathematics and Artificial Intelligence **49**(1-4) (April 2007) 101–115
17. Boley, M., Grosskreutz, H.: Non-redundant Subgroup Discovery Using a Closure System. In: Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases: Part I, Springer-Verlag (2009) 194
18. Bay, S., Pazzani, M.: Detecting group differences: Mining contrast sets. Data Mining and Knowledge Discovery **5**(3) (2001) 213–246
19. Bousquet, C. and Sadakhom, C. and Le Beller, C. and Jaulen, M.C. and Lillo-Le Louët, A.: Revue des signaux générés par une méthode automatisée sur 3324 cas de pharmacovigilance. Thérapie **61**(1) (2006) 39–47

Annexe C

La FCA pour la construction d'ontologie [BNT08b]

Formal Concept Analysis: A unified framework for building and refining ontologies

Rokia Bendaoud, Amedeo Napoli, and Yannick Toussaint

UMR 7503 LORIA, BP 239, 54506 Vandœuvre-lès-Nancy, FRANCE
{Rokia.Bendaoud,Amedeo.Napoli,Yannick.Toussaint}@loria.fr

Abstract. Building a domain ontology usually requires several resources of different types, e.g. thesaurus, object taxonomies, terminologies, databases, sets of documents, etc, where objects are described in terms of attributes and relations with other objects. One important and hard problem is to be able to combine and merge knowledge units extracted from these different resources within an homogeneous formal representation (such as a description logic or OWL). The purpose of this article is to show which kinds of resources should be available for designing a real-world ontology in a given application domain, and then how Formal Concept Analysis and its extension - Relational Concept Analysis- can be used for materializing an associated ontology. This resulting target ontology can then be encoded within OWL or a description logic formalism, allowing classification-based reasoning. A real-world example in microbiology is detailed. Finally, an evaluation including tests on recall and precision shows how source resources can be completed with other existing domain resources using a semi-automatic analysis process.

1 Introduction

Ontologies are the backbone of Semantic Web as they help software and human agents to communicate and to share domain knowledge [1]. In theory, an ontology is considered as an explicit specification of a domain conceptualization [15]. In practice, an ontology may depend on various resources with different types, e.g. thesaurus, vocabularies or dictionaries, sets of documents, databases. Moreover, the web makes an increasing number of ontologies available for reuse. None of these ontologies can pretend to be complete but rather brings a specific point of view on a particular domain. Besides ontologies, resources of other types exist but are heterogeneous and most of the time disconnected. One important need in the framework of semantic web is to take advantage of all these types of resources. Thus, there is a need for integrating or "pushing" these various types of resources with the objective of knowledge sharing, updating, dissemination, communication, and being complete as much as possible with respect to a given domain. However, this wide range of resources should be interoperable and resource contents are represented within a common standard language such as, e.g. OWL[1] (as assumed within the framework of semantic import of modular ontologies in [8]), but this is not always the case.

[1] OWL : Web Ontology Language

Following this way, this paper aims at presenting a framework based on Formal Concept Analysis (FCA) for integrating and preparing various types of resources for allowing collaboration, interoperability, and the design of a domain ontology for problem-solving and reasoning. This is the role of Formal Concept Analysis to fill the gap between resources of various types and a "target ontology" encoded within the OWL language. The paper introduces a framework where FCA and its extension, Relational Concept Analysis (RCA), can be considered as integrating processes for resource integration, leading from a set of heterogeneous resources to a set of formal and homogeneous ontologies, and finally to a given target ontology.

Three main types of resources are distinguished in the following: a thesaurus, a database, and a set of documents. In a standard way, the thesaurus provides a set of hierarchically organized classes. The database and the set of documents provide a set of pairs $(\texttt{object}, \texttt{attribute})$ (attribute or property) and a set of triples $(\texttt{object}_\texttt{i}, \texttt{relation}, \texttt{object}_\texttt{j})$. For example, the class of `firmicute` bacteria can be described by pairs, e.g. a set of attributes such as $\{\texttt{aerobic}, \texttt{negativeGram}, \texttt{spherical}\}$, and by triples such as the relation `ResistTo` whose co-domain includes ten families of antibiotics. These pairs and triples are taken into account within binary contexts for being processed by FCA and RCA for designing the final and integrated ontology. Moreover, pairs and triples will participate in defining a domain concept `C` with the help of necessary and sufficient conditions for testing the membership of an object `x` to the set of instances of `C`. As detailed later, pairs and triples also help in making explicit elements of implicit knowledge that are not directly accessible in the sole thesaurus. Meanwhile, the thesaurus –as understood here– plays a central role in the target ontology and for the domain expert (that is in charge of the interpretation of the extracted knowledge units for example). The thesaurus can be seen as a reference to which the target ontology can be compared. In this way, firstly, a class in the thesaurus may have been split into say two distinct classes, meaning that the attributes or relations observed in the other resources lead to the existence of these two classes: in this way, the elements of knowledge present in the original thesaurus have been made precise and completed. Secondly, two existing classes in the thesaurus may be merged into a more general class in the target ontology. An explanation may be that the two original classes in the thesaurus share a sufficient number of attribute for being identified in the new organization of the target ontology, meaning to some extent that the original distinction is not meaningful with respect to the resources examined during the process.

FCA and RCA are the processes on which is based the transformation between resources towards the target ontology. One important idea on which relies the process is the existence of a "source" or "pivot" ontology obtained from the thesaurus, and then to extend the source ontology by progressively adding units extracted from the resources under study. The addition of these units is based on the one hand on standard operations from FCA, such as apposition for example, but on the other hand on non standard operations such as RCA. Then, the

elements in the final concept lattice –built thanks to FCA– can be represented within the knowledge representation language OWL. In this way, FCA is considered as the "core" process in the design of the target ontology from a set of heterogeneous resources. This is the objective of this paper to explain how FCA and its extension RCA can be used for building, completing, and updating, a domain ontology. Firstly, FCA and RCA as well take into account all elements included within an ontology, namely objects (or individuals), attributes, and relations, for building concept lattices. Secondly, the FCA framework provides operations to manage concept lattices, e.g. updating the lattice when the set of objects or the set of attributes is modified, merging or linking concept lattices. Finally, the resulting concept lattices can be almost straightforwardly transformed into a concept hierarchy in a description logic (DL) $\mathcal{FLE}$ [5] or OWL concept hierarchy. A classifier can then be used for classification-based reasoning, e.g. answering queries. There are similar approaches but the novelty here lies in the articulation of the different operations for building up the target ontology. Moreover, an operational platform has been designed and detailed tests at the end of the paper show the capabilities of the approach and the efficiency of an FCA-based transformation approach.

The paper is organized as follows. The second Section discusses requirements for designing an ontology from a set of heterogeneous resources. The third Section introduces FCA and RCA, and the transformation process from a concept lattice to a concept hierarchy within a DL-based framework. The fourth Section presents a real-world example of the design of a target ontology from a set of heterogeneous resources in microbiology. An evaluation of the ontology design process follows. Related work is examined at the end of the paper.

2 Elements for building an ontology from heterogeneous resources

In this section, we analyze the basic objects and the associated resources that have to be considered for building an ontology in a given application domain. The domain chosen in this paper is microbiology, and two main kinds of basic objects are involved, i.e. bacteria and antibiotics. The problem is to build an ontology about resistance of bacteria to antibiotics on the base of a collection of heterogeneous resources. For bacteria, the following resources have been considered:

- The NCBI taxonomy (from the National Center for Biotechnology Information) includes 13380 species of bacteria,
- A collection of textual documents composed of 1244 abstracts has been selected by domain experts from PubMed (`http://www.ncbi.nlm.nih.gov/sites/entrez`), a large collection of texts in the NCBI library.
- The pathogenic bacteria database (`http://bac.hs.med.kyoto-u.ac.jp/`).

For antibiotics, a concept lattice of ligands has been designed based on expert available knowledge (involving mainly chemical properties of antibiotics).

2.1 Three main types of object descriptors

Ontologies are usually not built from scratch and several kinds of resources can be used. Actually, the type of the resources does not matter as much as the type of information the resources include. In this paper, three main types of object descriptors are distinguished, (OD1) hierarchical links, (OD2) binary attributes (or unary relations), and (OD3) relational attributes (or binary relations),

(OD1). In application domains, there are usually existing "source" hierarchies organizing domain objects, e.g. thesaurus, local ontologies from Swoogle [9]... Such hierarchies provide a global and structured view of the domain. In these hierarchies, a class denotes a set of objects and the relation between classes is set inclusion, while objects are "leaves" (terminal nodes): all objects in a class are also in the superclasses. For example, `Klebsiella-pneumoniae` (or `Klebsiella-P.`) is a kind of `Protebacteria`. Such classes can be compared to primitive concepts in description logics, as they do not have an explicit definition.

In the context of microbiology, the NCBI taxonomy has played the role of the "source" or the "pivot" domain hierarchy.

(OD2). For other kinds of resources, e.g. databases, domain objects are described by means of a set of attributes. For example, `helicobacter pylori` has the `negativeGram` attribute (in the pathogenic bacteria database). However, objects are not assigned to a class nor embedded into a hierarchy.

(OD3). Domain objects may be related to other objects. Such relations occur in texts, but not exclusively. For example, a sentence "We have previously reported that a significant percentage (44%) of *isoniazid-resistant Mycobacterium tuberculosis* strains carry an arginine to leucine mutation in codon 463 (R463L) in the catalase-peroxidase gene (katG)." indicates that there exists a *resistance* relation from *Mycobacterium tuberculosis* to *isoniazid*. Such type of relations has to participate to the definition of classes of objects as well as attributes.

The processing of textual resources. Given the texts and the databases listed above, attributes and relations between objects were extracted by the GATE system[2]. In the present framework, the use of GATE consists in two main operations. The first operation is an extraction of different entities of the domain, e.g. bacteria, antibiotics, etc. A second operation is the identification of relations existing in the analyzed texts, e.g. resistance, susceptibility, etc. For example, the analysis with respect to the resistance relation of a sentence such as "The genes conferring resistance to doxorubicin and daunorubicin in S. peucetius have been sequenced." Returns the tagged text shown below, used for describing a resistance relation between two objects, namely bacteria and antibiotics.

```
<Resistance>
   <Bacteria> S. peucetius </Bacteria>
```

[2] `http://gate.ac.uk/`

```
   <Antibiotic> doxorubicin </Antibiotic>
   <Antibiotic> daunorubicin </Antibiotic>
</Resistance>
```

2.2 From a reference ontology to a completed target ontology

The structure of the target ontology and its content has to take into account the three types of descriptors, *(OD1)*, *(OD2)* and *(OD3)* introduced here-above as hierarchical links, attributes, and relations respectively. Domain objects are grouped into the same class if and only if they share a given set of common attributes and relations. Both properties and relations are necessary and sufficient conditions for defining such a class of objects. For example, in microbiology, let us suppose that the `X` bacteria resists drug `D1`, the bacteria `Y` resists drug `D2`, and `D1` and `D2` are drugs of the family `D`. In this context, `X` and `Y` can be grouped in the same class as they share the relation "resisting drug from the class `D`". The resistance relation impacts on the definition of bacteria (here the domain of the relation). This shows in particular that attributes should be combined with relational attributes for forming richer and more precise definitions.

In the present framework, the NCBI taxonomy after being processed by FCA (to be explained after) has played the role of reference ontology *(OD1)*. The other resources that were analyzed to complete this reference ontology were describing genes, bacteria, and drugs.

The purposes of a target ontology depend partially on the type of queries one expects to ask. In the present context, the structure and the content of the target ontology should allow asking three main types of queries.

- *(Q1)*. Let o_1 and o_2 be two domain objects. Does there exist a class containing both objects or are these objects incompatible? What are the other objects in this class? How is this class defined ?
- *(Q2)*. Given a new object, say `x`, that has been observed with some attributes and relations with other objects. What is the best and the right way of inserting this object in the ontology? Is there a class already available for this object or a new class has to be created?
- *(Q3)*. What is the class of an object knowing the domain and/or the range of a relation? For example, when $r_1(o_1,o_2)$ and o_1 is an instance of $C_1 = \forall r_1.A_1$, then it can be inferred that o_2 is an instance of A_1.

3 Formal Concept Analysis

Formal Concept Analysis (FCA) and its extension Relational Concept Analysis (RCA) take into account the three main types of object descriptors discussed in Section 2. The FCA process builds concept lattices and provides various operations for managing concept lattices, and in particular merging sets of objects or sets of attributes. RCA extends the scope of FCA for dealing with relational attributes. Moreover, the resulting concept lattice can be transformed into a concept hierarchy represented within the description logic formalism to allow formal representation and reasoning.

3.1 Formal Concept Analysis

Formal concept analysis (FCA) [3] is a mathematical formalism allowing to derive a concept lattice from a formal context $\mathbb{K} = (G, M, I)$. FCA has been used for a number of purposes among which knowledge modeling, acquisition, and processing lattice and ontology design, information retrieval and data mining. In $\mathbb{K}$, G denotes a set of objects, M a set of attributes, and I a binary relation defined on the Cartesian product $G \times M$. In the binary table representing $I \subseteq G \times M$, the rows correspond to objects and the columns to attributes. The concept lattice is composed of *formal concepts* (or simply *concepts*) organized into a lattice by a partial ordering, i.e. a subsumption relation comparing concepts. A concept is a pair (A, B) where $A \subseteq G$, $B \subseteq M$, and A is the maximal set of objects sharing the whole set of attributes in B (and vice versa). In a concept (A, B), A is called the *extent* and B the *intent* of the concept. The concepts in a concept lattice are computed on the basis of a *Galois connection* defined by two derivation operators denoted by $'$:

$$A' := \{m \in M |\ gIm\ for\ all\ g \in A\}$$
$$B' := \{g \in G |\ gIm\ for\ all\ m \in B\}$$

A concept (A, B) verifies $A' = B$ and $B' = A$. The subsumption relation ($\sqsubseteq$) between a concept and a superconcept is defined as follows: $(A_1, B_1) \sqsubseteq (A_2, B_2) \Leftrightarrow A_1 \subseteq A_2 (or\ B_2 \subseteq B_1)$. Relying on this subsumption relation $\sqsubseteq$, the set of all concepts extracted from a context $\mathbb{K} = (G, M, I)$ is organized within a complete lattice, called *concept lattice* and denoted by $\mathfrak{B}(G, M, I)$.

The standard FCA process is able to deal with object descriptors of type *(OD1)* or *(OD2)*. Given a set of resources including such types of object descriptors, concept lattices provide a representation of the content of these resources. Then, the content of these resources can be merged using the FCA operation called *apposition*, as explained below.

	Proteobacteria	γproteobacteria	Actinobacteria	Bacilli
Helicobacter P.	X			
Klebsiella P.	X	X		
Mycobacterium S.			X	
Streptococcus P.				X
Klebsiella O.	X	X		

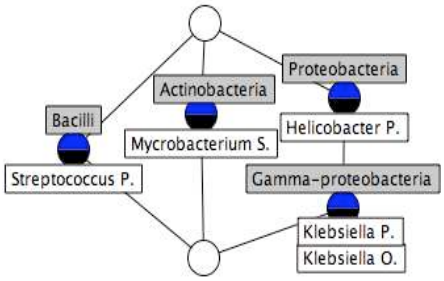


Fig. 1. The context Bacteria from the database NCBI $\mathbb{K}_1 := (G, M_1, I_1)$ and the associated concept lattice.

Building a lattice from a hierarchy (OD1 object descriptor). Transforming a set of objects organized within a hierarchy –or described by hierarchical links– into a lattice is a straightforward operation. The formal context $\mathbb{K}_1 := (G, M_1, I_1)$ is defined as follows: G is the set of domain objects, M_1 is the set of classes of objects organized into a hierarchy, and I_1 assigns to an object its class and all superclasses in the hierarchy. For example, the bacteria `Klebsiella P.` is classified in the NCBI hierarchical resource (in the domain of microbiology) as a γ`Proteobacteria`, which in turn is a subclass of `proteobacteria`. Figure 1 shows the context associated to NCBI classification and the corresponding concept lattice.

	spherical	sticks	negativeGram	positiveGram	aerobic	anaerobic
Helicobacter P.	×		×		×	
Klebsiella P.		×	×			×
Mycobacterium S.		×		×	×	
Streptococcus P.		×		×	×	
Klebsiella O.		×	×			×

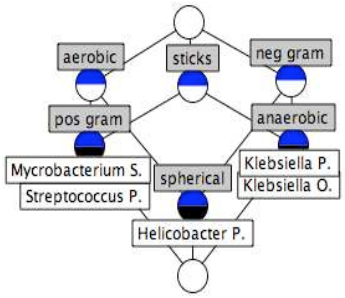


Fig. 2. The context Bacteria based on expert knowledge $\mathbb{K}_2 = (G, M_2, I_2)$ and the associated concept lattice.

Building a lattice from domain expert description of objects (OD2 object descriptor). A classification based on domain expert description of objects, i.e. involving *(OD2)* object descriptors, can be carried out as follows. A formal context $\mathbb{K}_2 := (G, M_2, I_2)$ is composed of a set G of objects, a set M_2 of attributes, and a relation $I_2 \subseteq G \times M_2$ where $I_2(g, m_2)$ states that g has the attribute m_2 (actually, the set G of objects is the same for context $\mathbb{K}_1$ and $\mathbb{K}_2$). Figure 2 shows an excerpt of such a context describing various bacteria, their attributes, and the corresponding concept lattice.

3.2 Apposition in FCA

At this point, there are two contexts $\mathbb{K}_1 := (G, M_1, I_1)$ and $\mathbb{K}_2 := (G, M_2, I_2)$, with the same set of objects G and two distinct sets of attributes, M_1 and M_2. There exists an operation in FCA for merging these two contexts into a single one called *apposition* [3].

Definition 1. *Let* $\mathbb{K}_1 = (G_1, M_1, I_1)$ *and* $\mathbb{K}_2 = (G_2, M_2, I_2)$ *be two formal contexts. When* $G = G_1 = G_2$ *and* $M_1 \cap M_2 = \emptyset$, $\mathbb{K} := \mathbb{K}_1 | \mathbb{K}_2 := (G, M_1 \cup M_2, I_1 \cup I_2)$ *is the* apposition *of the two contexts* $\mathbb{K}_1$ *and* $\mathbb{K}_2$.

The two contexts are $\mathbb{K}_1 = (G, M_1, I_1)$ shown in Figure 1 and $\mathbb{K}_2 = (G, M_2, I_2)$ shown in Figure 2. In the apposition context $\mathbb{K} = (G, M, I)$, G is the set of objects –the same set for $\mathbb{K}_1$ and $\mathbb{K}_2$– $M := M_1 \cup M_2$ where M_1 is the set of attributes in $\mathbb{K}_1$ –extracted from the NCBI hierarchy– and M_2 is the set of domain attributes in $\mathbb{K}_2$, and $I := I_1 \cup I_2$. The resulting concept lattice is presented in figure 3.

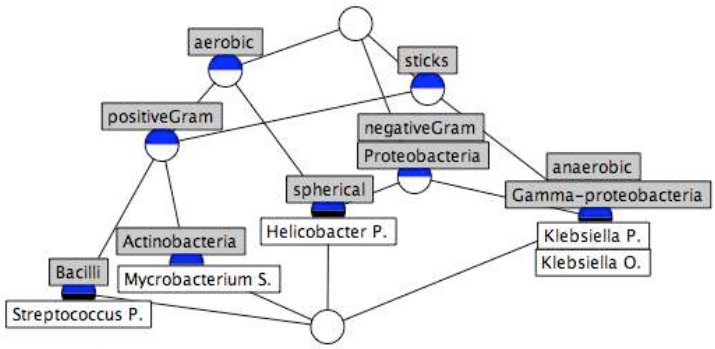


Fig. 3. The concept lattice resulting from the apposition of contexts $\mathbb{K}_1$ and $\mathbb{K}_2$.

3.3 Relational Concept Analysis

Relational Concept Analysis (RCA) [10] was introduced as an extension of FCA for taking into account relations between objects. In this way, a concept is described with standard binary attributes but also with relational attributes. A relational attribute, say r, describes the relation existing between objects that are instances of a concept, say c_1, the domain of the r relation, with objects that are instances of another concept, say c_2, the range of r relation. RCA has already been used in a previous work in text mining and ontology design [13].

More precisely, data in RCA are organized within a *relational context family* (RCF) composed of a set of contexts $\mathbb{K}_i = (G_i, M_i, I_i)$ and a set of relations $r_k \subseteq G_i \times G_j$. The sets G_i and G_j are the object sets of the contexts $\mathbb{K}_i$ and $\mathbb{K}_j$, called respectively the *domain* and the *range* of the relation r_k.

RCA uses the mechanism of *relational scaling* for defining the so-called relational attributes. For a relation, say $r : G_i \longrightarrow G_j$, linking objects from G_i to objects of G_j, a relational attribute is created and denoted by $r : c$, where C is concept in $\mathbb{K}_j$. Then, for an object $g \in G_i$, the relational attribute $r : c$ characterizes the "correlation" between g and $r(g) = h$ which is an instance of the concept $C = (X, Y)$ in $\mathbb{K}_j$. Some correlations can be considered such as the "existential correlation" –or existential scaling– where $r(g) \cap X \neq \emptyset$, and the "universal correlation" –or universal scaling– where $r(g) \subseteq X$. In the present work, only existential scaling is considered.

	FRB1	FRB3	ARB1	ARB2	HBA5	HBA10
Clarithromycin		×				×
Ciprofloxacin	×			×	×	
Cefotaxim		×	×			×
Macrolide			×			×

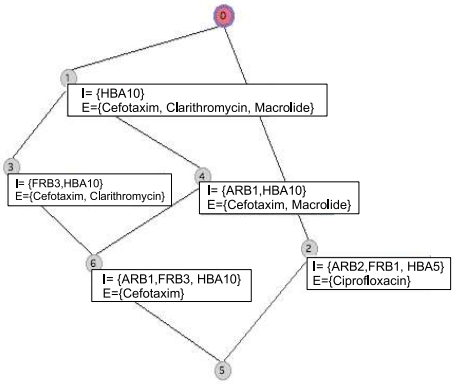


Fig. 4. The context Antibiotics $\mathbb{K}_3 = (G_3, M_3, I_3)$ and the associated concept lattice.

Table 1. The relation "ResistTo" between bacteria and antibiotics.

ResistTo	Clarithromycin	Ciprofloxacin	Cefotaxim	Macrolide
Helicobacter-P.		×		
Klebsiella-P.				×
Mycobacterium-S.			×	
Streptococcus-P.		×		
Klebsiella-O.	×			

Let us consider the relation between bacteria and antibiotics, where the first context is given by context apposition in Figure 3 and the second context $\mathbb{K}_3 = (G_3, M_3, I_3)$ is given in Figure 4. The relation `ResistTo` between bacteria and antibiotics is given in Table 1. The application of the RCA process based on the concept lattices of Figure 3 and Figure 4 produces the final concept lattice shown in Figure 5, where the relations explicitly computed by the RCA process are emphasized.

In more details, in Table 1, `Mycobacterium-S.` is related through `ResistTo` to `Cefotaxim` and `Streptococcus-P.` to `Ciprofloxacin`. Examining the lattice of antibiotics on Figure 4, it can be seen that `Cefotaxim` is in the extension of concepts `A0`, `A1`, `A3`, `A4`, and `A6`, while `Ciprofloxacin` is in the extension of concepts `A0` and `A2`. The relational attributes `ResistTo:A0`, `ResistTo:A1`, `ResistTo:A3`, `ResistTo:A4`, and `ResistTo:A6`, are associated to the object `Mycobacterium-S.`, while the relational attributes `ResistTo:A0` and `ResistTo:A2` are associated to `Streptococcus-P.` Then, a new concept lattice is built according to the extended context. At this point, as new concepts are been built, the

lattice construction process is iterated and new relational attributes are associated to the bacteria objects whenever possible. If this is the case, the RCA process is iterated again. If this is not the case, this means that the fix-point of the RCA process has been reached and that the final concept lattice has been obtained. This final lattice is given on Figure 5 (lattice on the left). In particular, it can be seen that `Mycobacterium-S.` is in the extension of concept `C13` while `Streptococcus-P.` is in the extension of `C12`. All relational attributes introduced above are respectively associated to `C13` and `C12` (attributes are inherited from upper concepts). It can be noticed that the identity of a concept is constant during the whole RCA process.

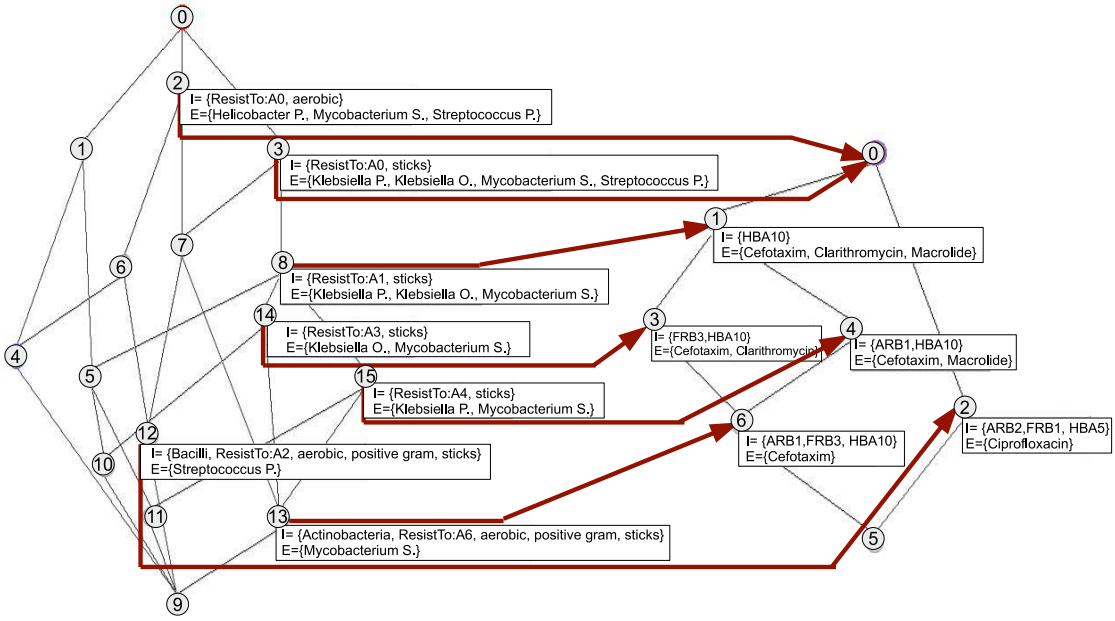


Fig. 5. The lattice resulting from the RCA process applied to object descriptors of type *(OD3)*.

4 From concept lattice to DL formalism

4.1 The representation of formal concepts into $\mathcal{FLE}$ concepts

The transformation of the final concept lattice resulting from RCA is based on a transformation, called τ, into a DL knowledge base (KB). The τ transformation allows to introduce primitive and defined concepts, and thus to apply a DL-based

reasoner for problem-solving and complex query answering. The target DL formalism is $\mathcal{FLE}$ [5], that includes the constructors $\top$ (top), $\bot$ (bottom), $C \sqcap D$ (concept conjunction), $\forall r.C$ and $\exists r.C$ (universal and existential role quantifications). This set of constructors is large enough for representing all elements from the final concept lattice. The profile of the τ transformation is the following:

$\tau : \underline{\mathfrak{B}}(G_f, M_f, I_f) \longrightarrow TBox \cup ABox$, where $\underline{\mathfrak{B}}(G_f, M_f, I_f)$ is the final concept lattice, TBox and ABox being the DL components on which the target ontology will be based. The concept lattice $\underline{\mathfrak{B}}(G_f, M_f, I_f)$ results from the FCA operations applied to the reference lattice –mainly appositions and relational scalings– which here is the concept lattice associated to the NCBI hierarchy (Figure 1). More precisely, the τ transformation works as follows:

- An attribute $m_1 \in M_1$, where $\underline{\mathfrak{B}}(G, M_1, I_1)$ of the concept lattice associated to the NCBI hierarchy, is transformed into an atomic –or primitive– concept in the TBox. This means that a class in the NCBI hierarchy is represented as an atomic concept, e.g. τ`(Proteobacteria)` = `Proteobacteria`.
- An attribute in a context distinct from $\underline{\mathfrak{B}}(G, M_2, I_2)$, e.g. in the Bacteria context associated to expert knowledge (see Figure 2), is transformed as a conceptual expression of the form $\exists$`m.`$\top$. For example, τ`(negativeGram)` = $\exists$`negativeGram.`$\top$.
- A relational attribute $r \in R$ is transformed in the TBox as an atomic role $\tau(r)$, e.g. τ`(ResistTo)` = `ResistTo`, i.e. an atomic role with the same name in the TBox.
- A formal concept $C = (X, Y)$ is transformed in the TBox as a defined concept formed by the conjunction of primitive concepts and existential role quantifications. For example, $\mathtt{C}_{12}$ = `Bacilli` $\sqcap$ $\exists$`sticks.`$\top$ $\sqcap$ $\exists$`aerobic.`$\top$ $\sqcap$ $\exists$`positiveGram.`$\top$ $\sqcap$ $\exists$`ResistTo.`$\mathtt{A}_2$ (see Figure 5).
  A subsumption relation between concepts is transformed as a general concept inclusion: the $C_1 \sqsubseteq C_2$ subsumption relation in the lattice becomes $\tau(C_1) \sqsubseteq \tau(C_2)$.
- An object $g \in G$ is transformed as an individual $\tau(g)$ in the ABox, e.g. `Staphylococcus aureus` becomes the individual τ`(Staphylococcus-aureus)`.

Here are some examples of defined concepts:
$\mathtt{C}_2$ = $\exists$`aerobic.`$\top$ $\sqcap$ $\exists$`ResistTo.`$\mathtt{A}_2$
$\mathtt{C}_{12}$ = `Bacilli` $\sqcap$ $\exists$`sticks.`$\top$ $\sqcap$ $\exists$`aerobic.`$\top$ $\sqcap$ $\exists$`positiveGram.`$\top$ $\sqcap$ $\exists$`ResistTo.`$\mathtt{A}_2$
$\mathtt{C}_{13}$ = `Actinobacteria` $\sqcap$ $\exists$`sticks.`$\top$ $\sqcap$ $\exists$`aerobic.`$\top$ $\sqcap$ $\exists$`positiveGram.`$\top$
$\sqcap$ $\exists$`RestTo.`$\mathtt{A}_6$

4.2 Reasoning within the DL formalism

The main reasoning operations that can be drawn are concept instantiation and concept subsumption, e.g. detecting the class of an individual –class stands here for concept extent or in DL terms as the set of instances of a concept– analyzing the range of a relation, or comparing concepts. Details for each reasoning operation are given below in the context of the microbiology example.

Instantiation consists in finding the class of an object (or individual). Let o_1 be an object with attributes {a,b} and relational attributes {r_1.A_1,r_2.A_2}, and belonging to classes {C_3,C_4} in the NCBI hierarchy. Then, the class of o_1 is the most general class X in the target ontology such that: $X \sqsubseteq C_3 \sqcap C_4 \sqcap \exists a.\top \sqcap \exists b.\top \sqcap \exists r_1.A_1 \sqcap \exists r_2.A_2$. This is a way of answering a question such as "What is the class of the object Streptococcus pneumoniae", whose attributes are {aerobic,positiveGram,sticks}, relational attributes are {ResistTo:A_2}, and belonging to the class {Bacilli} in NCBI. According to the final lattice given in Figure 5, the answer is the concept C_{12}.

A second task consists in determining whether two objects o_1 and o_2 have the same class. A simple way is to find the class of o_1, then the class of o_2, and then to test whether the two classes are equivalent. For example, let us consider the objects Klebsiella-O. and Streptococcus-P.. Klebsiella-O. is an instance of the class C_{14} and Streptococcus-P. is an instance of the class C_{12} (see Figure 5). In this case, the fact that $C_{14} \sqcap C_{12} = \bot$ implies that both objects do not belong to the same class (or are incompatible).

Finally, the third task consists in detecting the class of an object knowing the domain or the range of a relation. Let us consider the instantiated relation $r_1(o_1,o_2)$. When o_1 is an instance of the class $C_1 = \forall r_1.A_1$, it can be inferred that o_2 is an instance of A_1. When o_2 is an instance of A_1, it can be inferred that o_1 is an instance of a class defined by an expression of the form either $\forall r_1.A_1$ or $\exists r_1.A_1$. For example, knowing that ResistTo(b_1,a_1) with a_1 as an instance of Ciprofloxacin, it can be inferred that b_1 is a bacteria, instance of concept C_{12} in Figure 5 (Streptococcus-P.).

5 Evaluation

There is no absolute and objective criteria to evaluate an ontology. Thus we decided to compare the target ontology to the NCBI thesaurus considered as the reference ontology.

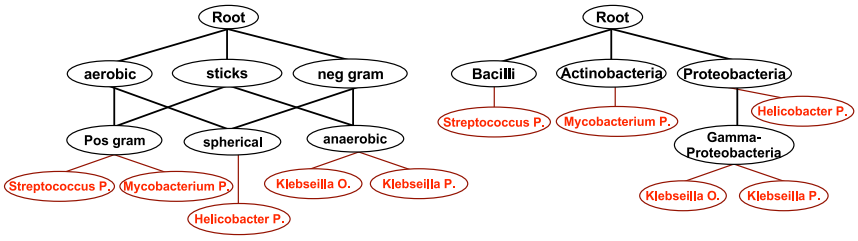


Fig. 6. Example of two ontologies, Ω_{target}(left) and $\Omega_{reference}$(right)

We followed Maedche and Staab [1] approach, adapted by Cimiano et al. [12]. However, the representation of the lattice into DL following the function α makes

the presentation of the evaluation quite different. This evaluation relies on similarity between sets of instances. First, for any class of the target ontology, its closest class in the reference ontology is computed.

Computing the closest class. The closest class is computed using the Euclidian distance on the set of instances. Let G be the set of objects, Ω_1 and Ω_2 be the two ontologies. For each class $C_1 \in \Omega_1$, and for each class $C_2 \in \Omega_2$, vectors V_{C_1} and V_{C_2} are defined as: $\forall g_k \in G$: if g_k is an instance of C_i then $V_{C_i}[g_k] = 1$ else $V_{C_i}[g_k] = 0$. Then:

$$Distance(V_{C_1}, V_{C_2}) = (\sum_{k=0}^{|G|}(V_{C_1}[g_k] - V_{C_2}[g_k])^2)^{1/2}$$

For the examples from figure 6: $Distance(\texttt{sticks}, \texttt{proteoacteria}) = \sqrt{3}$, and C_1 is the closest class of C_2 iff $\forall C \in \Omega_1 - \{C_1\}$ with $Distance(V_C, V_{C_2}) \geqslant (V_{C_1}, V_{C_2})$.

Computing precision and recall. We introduce three measures for ontology comparison. The precision for a given class $C_1 \in \Omega_1$ is computed with its closest class $C_2 \in \Omega_2$ as the proportion of instances from C_1 common to C_2. Recall is the proportion of instances from C_2 common to C_1.

$$Precision(C_1) = \frac{|C_1 \cap C_2|}{|C_1|}, \quad Recall(C_1) = \frac{|C_1 \cap C_2|}{|C_2|}$$

Precision (resp. recall) is the average of precision (resp. recall) on all classes from the target ontology. F-measure is also defined as the harmonic mean of precision and recall. Let N be the number of classes in the target ontology and C_i a class in Ω_1:

$$P(\Omega_1, \Omega_2) = \frac{\sum_{i=1..N}(Precision(C_i))}{N}, \quad R(\Omega_1, \Omega_2) = \frac{\sum_{i=1..N}(Recall(C_i))}{N}$$

$$F(\Omega_1, \Omega_2) = \frac{2 * P(\Omega_1, \Omega_2).R(\Omega_1, \Omega_2)}{P(\Omega_1, \Omega_2) + R(\Omega_1, \Omega_2)}$$

For examples from figure 6, we have:
$P = \frac{1+1+1+1+\frac{1}{2}+1+1}{7} = 92,85\%$
$R = \frac{1+1+\frac{1}{3}+\frac{4}{5}+\frac{3}{5}+1}{7} = 81,90\%$

We have worked on two separate experiments: the first using only FCA (attributes) and the second using FCA + RCA (attributes + relational attributes). Table 2 presents the resulting precision, recall, and F-measure. This table presents also the number of classes with 100% precision and recall which FCA (and/or RCA) defines, i.e. gives necessary and sufficient conditions. RCA defines more classes than FCA (see Table 2). In following, we present an example of classes defined just with FCA and an example of classes defined by RCA.

Table 2. The results of the evaluation of the FCA and RCA

	Number of classes	Number of properties	Number of relation	Precision	Recall	F-Measure	Defined concepts
FCA	58	13	0	76,52%	81,14%	78,76%	8/19
RCA	152	13	55	66,88%	82,07%	73,70%	12/19

FCA shows better precision and recall (see Table 2) than RCA. However, some classes need relations in their definitions, explaining why RCA shows a better number of defined classes. For example, let us consider the class C_9, i.e. `({Neisseria gonorrhoeae}, {Betaproteobacteria, Neisseria, Proteobacteria})`. Using FCA does not allow to find a closest class with precision and recall of 100%. Insted, using RCA, allows to build the class C_{150} that has exactly the same set of instances `({Neisseria gonorrhoeae}`, with precision and recall of 100%.

6 Related work

In [4], the authors present a cooperative machine learning system called ASIUM, which is able to acquire semantic knowledge from syntactic parsing. The system ASIUM successively aggregates clusters to form new concepts and the hierarchies of concepts from the ontology. The ASIUM approach differs from our approach because the former is not based on the same classification approach, does not work with heterogeneous resources, and does not try to complete ontologies for building a target ontology. In [12], the authors use an approach similar to the preceding approach, but they use the FCA for building the concept hierarchy. Regarding the work of Cimiano et al., our approach involves the use of FCA and RCA as well as taking into account heterogeneous resources.

The extraction of relational attributes allows a better definition of concepts. In [7], the authors propose the system "Lexical Navigation" for extracting the (non hierarchical) relations. Their idea is to use a lexical network containing domain-specific vocabularies and relationships that are automatically extracted from a collection of documents. In the same way, the work in [11] proposes to use a learning method to extract syntactic patterns. This method extracts manually relations between terms from texts and searches to generalize the terms and the relation between these terms. In comparison to our method, the two preceding methods try to cluster terms with a different classification approach, taking into account relations in texts but without a systematic approach as FCA and RCA. The facts of dealing with heterogeneous resources and with a final target ontology are also different.

Another approach described in [2] consists in extracting association rules from a collection of texts and keeping the rules having a given support and frequency. The objectives could be compared but the classification methods used in this work and ours are quite different.

In [6], the authors propose to merge two ontologies for building a new one. The proposed method takes as input a set of documents. NLP techniques are used to capture two formal contexts encoding the relationships between documents and concepts in each ontology. This method combines the knowledge of the collection of texts and expert knowledge. Comparing with our approach, the approach of Stumme et al. uses the texts for merging and not for enriching the two ontologies. The authors in [14] propose to enrich an existing ontology using on-line glossaries. They use natural language definitions of each class and convert them into formal definitions (OWL), compliant with the core ontology property specifications. Then, this method needs an existent core ontology for adding the transverse relations that is not our case.

7 Conclusion

In this paper, we have presented an original approach for building a target domain ontology in considering resources of different types, such as a thesaurus, term hierarchies, databases, and sets of documents. In these resources, objects are described in terms of attributes and relations with other objects. Using the FCA process and its extension RCA, these different resources can be represented as concept lattices. These concept lattices are used to complete a chosen reference concept lattice, that will be the basis of the target ontology. Then, this final concept lattice is transformed within a description logic formalism. Complex question-answering and classification-based reasoning can then be carried out using the classifier in the framework of description logics. A real-world example in microbiology has been detailed, showing that the approach is fully operational.

In this paper, only a part of the available and potential knowledge implicitly lying in the different resources has been extracted for analyzing the phenomenon of bacteria resisting to antibiotics. In future work, we plan to extend the target ontology by extracting other objects that are of importance, e.g. genes, codons, etc., and, as well, other relations, which include composition and spatial relations between bacteria, genes, and other biological actors present in the texts. In this way, a more precise representation of the content of texts will be designed and used for characterizing texts in microbiology. Finally, such a characterization could be used for comparing, classifying, and computing similarities between texts on the basis of their contents. This could also lead to sophisticated kinds of reasoning on texts, e.g. case-based reasoning and adaptation of a texts.

References

1. Maedche A. *Ontologies Learning for the Semantic Web.* Springer, 2002.
2. Maedche A. and Staab S. Discovering conceptual relation from text. In *14th European Conference on Artifical Intelligence (ECAI'00)*, pages 321–325, Berlin, Germany, 2000.
3. Ganter B. and Wille R. *Formal Concept Analysis, Mathematical Foundations.* Springer, 1999.

4. Faure D. and Nedellec C. A corpus-based conceptual clustering method for verb frames and ontology acquisition. In *Workshop on Adapting lexical and corpus resources to sublanguages and applications (LREC'98)*, 1998.
5. Baader F., Calvanese D., McGuinness D.L., Nardi D., and Patel-Schneider P.F., editors. *The description logic handbook: theory, implementation, and applications*. Cambridge University Press, New York, NY, USA, 2003.
6. Stumme G. and Maedche A. Fca-merge: Bottom-up merging of ontologies. In *International Joint Conference on Artificial Intelligence (IJCAI'01)*, pages 225–234, 2001.
7. Cooper J.W. and Byrd R.J. Lexical navigation: Visually prompted query expansion and refinement. In *2nd International Conference on Digital Libraries (DL'97)*, pages 237–246, 1997.
8. Pan J.Z., Serafini L., and Zhao Y. Semantic import: An approach for partial ontology reuse. In *1st International Workshop on Modular Ontologies (WoMO'06) In ISWC 2006*, 2006.
9. Ding L., Finin T.W., Joshi A., Pan R., Scott Cost R., Peng Y., Reddivari P., Doshi V., and Sachs J. Swoogle: a search and metadata engine for the semantic web. In Grossman D., Gravano L., Zhai C., Herzog O., and Evans D.A., editors, *International Conference on Information and Knowledge Management (CIKM'04)*, pages 652–659. ACM, 2004.
10. Rouane-Hacene M., Huchard M., Napoli A., and Valtchev P. A proposal for combining formal concept analysis and description logics for mining relational data. In Kuznetsov S.O and Schmidt S., editors, *Proceedings of the 5th International Conference on Formal Concept Analysis (ICFCA 2007), Clermont-Ferrand*, LNAI 4390, pages 51–65. Springer, Berlin, 2007.
11. Aussenac-Gilles N., Biébow B., and Szulman S. Revisiting ontology design: A method based on corpus analysis. In Dieng R. and O. Corby, editors, *12th Internationnal Conference in Knowlegde Engineering and knowledge Management (EKAW'00)*, volume 1937, pages 172–188, 2000.
12. Cimiano P., Hotho A., and Staab S. Learning concept hierarchies from text corpora using formal concept analysis. *Journal of Artificial Intelligence Research (JAIR)*, 24:305–339, 2005.
13. Bendaoud R., Rouane-Hacene M., Toussaint Y., Delecroix B., and Napoli A. Text-based ontology construction using relational concept analysis. In Flouris G. and d'Aquin M., editors, *Proceedings of the International Workshop on Ontology Dynamics, Innsbruck (Austria)*, pages 55–68, 2007.
14. Navigli R. and Velardi P. Ontology enrichment through automatic semantic annotation of on-line glossaries. In Staab S. and Svátek V., editors, *15th Internationnal Conference in Knowlegde Engineering and knowledge Management (EKAW'06)*, volume 4248, pages 126–140, Podebrady, Czech Republic, 2006. Springer.
15. Gruber T.R. Toward principales for the design of ontologies used for knowledge sharing. In Guarino N. and R. Poli, editors, *Formal Analysis in Conceptual Analysis and Knowledge Representation*, The Netherlands, 1993. Kluwer Academic.