Proposition d'une mesure de voisinage entre textes : Application à la veille stratégique

Annette Casagrande

UNIVERSITÉ DE GRENOBLE

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Mathématiques-Informatique**

Arrêté ministériel : 7 août 2006

Présentée par

Annette CASAGRANDE

Thèse dirigée par **Laurent VUILLON**
et codirigée par **Humbert LESCA**

préparée au sein **des laboratoires LAMA et CERAG**
et de **l'école doctorale MSTII**

Proposition d'une mesure de voisinage entre textes : Application à la veille stratégique

Thèse soutenue publiquement le **3 juillet 2012**,
devant le jury composé de :

Monsieur, Matthieu LATAPY
Directeur de Recherche, CNRS, Rapporteur

Monsieur, Michel LÉONARD
Professeur, Université de Genève, Rapporteur

Monsieur, Éric GAUSSIER
Professeur, Université Joseph Fourier, Examinateur

Monsieur, Yves LECHEVALLIER
Directeur de Recherche, INRIA, Examinateur

Monsieur, Laurent VUILLON
Professeur, Université de Savoie, Directeur de thèse

Monsieur, Humbert LESCA
Professeur, Université Pierre Mendès France, Co-Directeur de thèse

« Ce n'est qu'en essayant continuellement que l'on finit par réussir.... En d'autres termes... Plus ça rate et plus on a de chances que ça marche... »

« S'il n'y pas de solution, c'est qu'il n'y a pas de problème »
Proverbes Shadoks

Remerciement

Je tiens à remercier tous ceux qui d'une façon ou d'une autre ont contribué à ce travail de thèse.

Je remercie Matthieu Latapy et Michel Léonard qui ont accepté d'être mes rapporteurs, Eric Gaussier et Yves Lechevallier qui ont accepté de faire partie de mon jury.

Un grand merci à Laurent Vuillon et Humbert Lesca de m'avoir encadrée pendant ces quatres années.
Je remercie Laurent Vuillon d'avoir accepté tout de suite ce projet de thèse et de m'avoir fait confiance. Merci pour tes conseils, tes encouragements, ton optimisme (qui m'a beaucoup aidé dans les moments difficiles), ta gentillesse.
Je remercie Humbert Lesca de m'avoir fait découvrir la veille stratégique et de m'avoir emmenée sur le terrain. Merci pour votre gentillesse, vos conseils, vos histoires et anecdotes.

Je remercie également tous les membres du LAMA qui m'ont toujours bien accueillie et plus particulièrement : Julien Olivier, Matthieu Simonet, Matthieu Bonnivard, Mehmet Ersoy, Timack Ngom, Florian Hatat, Pierre-Etienne Meunier, Marguerite Gisclon, Céline Labart, Céline Acary-Robert, Christophe Raffalli, Pierre Hyvernat,

Guillaume Theyssier, Tom Hirschowitz.

A un grand merci à tous mes relecteurs pour leurs différentes contributions à ce mémoire : Julien Olivier, Cristina Breuil, Alex Buitrago, Céline Labart, Thierry Casagrande, Joëlle Casagrande, Cyrille Jost, Laure Daudin, Marie-Laurence Caron-Fasan, Fanny Vallet, Jean-Charles Marty, Cyrille Jost, Bénédicte Branchet et Sarah Setton.

Merci à Jean-Charles Marty, Thibault Carron, Fanny Vallet, Olivier Desrichard et Yolaine Cultiaux pour leur collaboration.

Merci à toute ma famille, en particulier mes parents, ma sœur, mon frère de m'avoir soutenue pendant ces 4 années de thèse.

Merci Cyrille pour ton amour, ta présence, tes encouragements, ta patience, nos discussions, tes conseils, ...

Table des matières

Introduction générale

Si la surinformation n'est pas un phénomène nouveau [SJ12, Gui12], elle prend une ampleur grandissante du fait de l'utilisation toujours plus importante des nouvelles technologies de l'information dans l'entreprise. Les études montrent que le volume des données générées au sein des entreprises croît de 50% par an. En outre, les informations les plus difficiles à gérer comme les flux texte, audio et vidéo non structurés représentent 80 % de ces données. Ceci n'est pas sans conséquences pour les salariés qui sont confrontés à un triple défi [CKI07] :

- l'accélération de la prise de décision. La perception par les salariés de l'exigence de prendre des décisions dans un laps de temps plus court est passée de 65.5% en 2001 à 70.9% en 2005.
- l'augmentation de la surcharge informationnelle. En 2005 les salariés étaient 79.3% à considérer que le volume d'informations à traiter était trop important alors qu'ils étaient 71.7% en 2001.
- l'augmentation de la surcharge cognitive. La proportion de salariés estimant « passer davantage de temps à classer l'information » est passée de 42.9% en 2001 à 57.2% en 2005.

Or « *la saturation d'informations conduit d'abord à la dégradation du processus de décision. Les recherches montrent qu'il existe un nombre optimal d'informations à recueillir pour prendre une décision. Au-delà d'une certaine quantité d'information, la qualité du processus décisionnel baisse, tant d'un point de vue de la qualité (décision rationnelle dans le contexte), que du temps pour prendre la décision (une*

décision qui intervient trop tard n'est pas bonne) »[SR10].

Dans ce contexte, la veille stratégique a pour objet l'amélioration du processus décisionnel des entreprises. Cette discipline est basée sur l'observation et l'analyse de l'environnement scientifique, technique, technologique et les impacts économiques présents et futurs pour en déduire les menaces et les opportunités de développement[age]. L'équipe de veille stratégique du CERAG [1] a proposé la méthode VASIC pour « permettre à l'entreprise d'agir vite, au bon moment, avec le maximum d'efficacité et le minimum de ressources, dans le but de contribuer à sa compétitivité durable »[Les03]. Pourtant, cette méthode, pertinente dans le cas d'informations préalablement sélectionnées, est mise en échec de par l'absence d'outils de gestion de la surinformation.

Ce manque d'outils se ressent dans la préparation de la séance de création collective de sens qui est au centre de la méthode VASIC. On entend par création collective de sens : « *un groupe de personnes [qui] accepte volontairement de mettre en commun (en collectif) leurs capacités de détecter des événements, d'en parler, de les interpréter ensemble et d'en tirer des enseignements utiles pour l'action* »[Les03]. Pour cela, il est nécessaire de sélectionner parmi les nombreuses informations collectées par l'entreprise des informations entre lesquelles il est possible d'établir des liens. On parlera alors d'« **informations voisines** ». Face à la surcharge d'information, comment sélectionner des informations voisines ? Comment mesurer la proximité entre deux informations ? Nous porterons notre attention sur les informations textuelles.

Dans le chapitre I, nous définirons les différents types d'informations utiles à l'entreprise dont les informations anticipatives. Nous présenterons ensuite le processus VASIC qui aide à la collecte et l'interprétation des informations anticipatives. Nous terminerons ce chapitre en expliquant la problématique rencontrée par les entreprises lors de la mise en place du processus VASIC et plus précisément lors de la séance de création collective de sens avec la sélection des informations voisines.
Afin de répondre à la demande de la veille stratégique, nous présenterons dans le chapitre II les techniques existantes en Recherche d'Information. Il existe deux grandes

1. Centre d'Etudes et de Recherches Appliquées à la Gestion

approches pour obtenir des informations : la recherche de documents (à l'aide de mots-clefs) ou la navigation dans des documents. Nous proposons une grille de lecture des outils de recherche ou de navigation en 5 fonctionnalités :

- l'abstraction : rendre un texte compréhensible par l'ordinateur,
- le filtrage : suppression des éléments jugés inutiles,
- le lissage : adaptation de l'abstraction pour améliorer les résultats
- la mesure : calcul évaluant la proximité entre deux textes ou une requête et un texte,
- la visualisation des résultats.

Dans le chapitre III, nous définirons notre mesure de voisinage en nous appuyant sur la définition des informations voisines. Cette mesure prend en compte les mots communs, les synonymes et les mot cooccurrents. Nous expliquerons la construction de graphes à partir du calcul des mesures de voisinages entre les textes d'une base. Nous comparerons nos résultats avec ceux obtenus avec une mesure classique en recherche d'information, la mesure cosinus. Nous démontrerons les propriétés observées graphiquement et le comportement des mesures (mesure de voisinage et cosinus) sur des textes aléatoires.
Cette mesure de voisinage a été implémentée dans un prototype nommé Alhena. Ce prototype calcule à partir des textes les mesures de voisinages et trace les graphes liés à ces mesures. Dans le chapitre IV, nous présenterons le prototype Alhena du côté utilisateur (présentation des différents écrans) et du côté informatique (algorithme et complexité).
Dans le dernier chapitre de cette thèse (chapitre V), nous montrerons l'utilité du prototype Alhena en veille stratégique au travers de deux expériences :

- la première a été menée sur le thème de la valorisation du CO_2,
- le seconde sur le bioéthanol montrera un apport particulier d'Alhena dans le cadre d'un dispositif de veille : l'élargissement du champ de vision (« peripheral vision »).

Dans un deuxième temps, nous validerons la pertinence de notre mesure en comparant les résultats obtenus avec Alhena et le classement « humain » d'une base d'articles de Reuters.
Nous montrerons que la mesure de voisinage peut aussi être utilisée dans d'autres

domaines :

- application en littérature : nous confronterons notre mesure aux Exercices de Styles de Raymond Queneau.
- application en psychologie : nous présenterons le travail réalisé en collaboration avec des chercheurs en psychologie sur l'analyse de situations de mémoire.
- application en informatique : nous décrirons notre collaboration avec des chercheurs en informatique sur l'analyse de tchats.

Enfin nous conclurons et proposerons différentes pistes de recherches.

Références bibliographiques

[age] AgentIntelligent.com - veille strategique - définition et objectifs. http ://www.agentintelligent.com/veille/veille_strategique.html. 2

[CKI07] E. Campoy, M. Kalika et H. Isaac : Surcharge informationnelle, urgence et tic : l'effet temporel des technologies de l'information. *Management et Avenir*, 13:153, 2007. 1

[Gui12] H. Guillaud : Notre surcharge informationnelle en perspective. http ://www.internetactu.net/2012/02/29/lift12-notre-surcharge-informationnelle-en-perspective/, 2012. 1

[Les03] H. Lesca : *Veille stratégique : la méthode L.E.SCAnning*. Gestion en liberté, ISSN 1625-3132. EEd. EMS, Colombelles, 2003. 2

[SJ12] A. Saint-Jude : From gutenberg to zuckerberg, information overload and social networks, 2012. 1

[SR10] C. Sauvajol-Rialland : La surcharge informationnelle dans l'organisation : les cadres au bord de la « crise de nerf ». *Le magazine de la communication de crise et sensible*, 19, 2010. 2

Chapitre I

La veille stratégique

Ce chapitre est consacré à la présentation de la veille stratégique et la problématique rencontrée lors de la mise en place d'un dispositif de veille. La veille stratégique est définie comme « *une activité continue et en grande partie itérative visant à une surveillance active de l'environnement technologique, commercial, ..., pour anticiper les évolutions* »(AFNOR). Cette activité nécessite la collecte d'informations. Il existe divers types d'informations utiles aux entreprises ; cependant seules les informations anticipatives sont utiles pour réaliser une veille stratégique efficace. L'équipe de veille stratégique du CERAG a proposé le processus VASIC (veille anticipative stratégique et intelligence collective) qui aide à la collecte et l'interprétation des informations anticipatives. La problématique, rencontrée par les entreprises lors de la mise en place du processus VASIC, concerne la préparation de la séance de création collective de sens (groupe de personnes qui acceptent de manière volontaire de partager leurs capacités à détecter des évènements, à en parler, à les interpréter et à en tirer des enseignements utiles à l'action). Lors de cette préparation, des informations, entre lesquelles des liens peuvent être établis (« informations voisines »), doivent être sélectionnées : face à la surcharge d'information, comment sélectionner des informations voisines ?

Sommaire

L'Association Française de NOrmalisation (AFNOR) [AFN98] définit la veille comme « *une activité continue et en grande partie itérative visant à une surveillance active de l'environnement technologique, commercial, ..., pour anticiper les évolutions* ». Dans cette veine, l'équipe de veille stratégique du CERAG [1] travaille depuis de nombreuses années sur la veille et plus précisément sur la veille anticipative stratégique et intelligence collective (VASIC). On peut en donner la définition suivante :« *la veille anticipative stratégique est un processus considéré comme un système d'information particulier tourné vers l'extérieur et vers le futur de l'entreprise (ou autre organisation). Elle permet à l'entreprise d'acquérir des informations destinées à nourrir la réflexion et la prise de décisions stratégiques et accroître sa réactivité* »[ML10].

Les travaux de recherche de l'équipe de veille stratégique ont démarré dès le début des années 1980 notamment sur invitation du professeur Thiétart. À sa demande un ouvrage a été rédigé, intitulé : « Système d'information pour le management stratégique, l'entreprise intelligente »[Les86]. Cet ouvrage a été rédigé notamment sur la base d'entretiens avec plusieurs dizaines de dirigeants d'entreprises françaises.

Ainsi est-il apparu très tôt que les responsables d'entreprises étaient demandeurs de concepts clairs concernant la veille stratégique, mais plus encore de méthodes de travail « pour faire » la veille stratégique. C'est pourquoi les travaux de l'équipe ont été orientés, dès le début :

- vers la production de connaissances actionnables [Arg96] : « *Actionable knowledge is any type of explicit or tacit knowledge that can be formalized, in order to allow its use in a causal manner by an entity (e.g., group of experts, system, user, knowledge base) or a service.* »[YLBH10].
- et la recherche-intervention sur le terrain : « *la recherche intervention en sciences de gestion entend produire des connaissances à la fois scientifiques et utiles à l'action* »[Dav00].

En outre, les dirigeants interviewés ont exprimé un intérêt très faible pour les informations de veille collectées et communiquées par leur service de documentation. Les expressions qui revenaient le plus souvent étaient : « je suis submergé par les informations que l'on me communique et que je n'ai pas le temps de lire ; elles ne me servent à rien ! » ou encore « je n'ai pas besoin de cimetière d'information ! ». Le

1. Centre d'études et de recherches appliquées à la gestion de Grenoble

besoin ressenti était plutôt d'extraire la quintessence des données recueillies par le service de documentation. D'où les questions : de quelle information les entreprises ont-elles besoin (section 1) ? comment tirer du sens, utile pour prendre des décisions, à partir des informations après les avoir collectées et sélectionnées (section 2) ?

Fruit de leurs nombreuses recherches, le modèle VASIC a été proposé par l'équipe de veille stratégique du CERAG ; il a comme finalité « *de permettre à l'entreprise d'agir vite, au bon moment, avec le maximum d'efficacité et le minimum de ressources* »[Les03a]. Ce modèle sera détaillé dans la section 2. La section 3 présentera la séance de création collective de sens qui est au centre de la méthode VASIC.
Enfin nous terminons par la présentation du concept d'informations voisines utiles pour optimiser le processus VASIC (section 4). Nous précisons son utilité dans le processus VASIC, face au problème de la surcharge d'information [EM04] notamment occasionnée par l'usage de l'Internet.

1 De quelle information les entreprises ont-elles besoin ?

1.1 Qu'est-ce qu'une information ?

Il existe de nombreuses définitions du terme information qui dépendent du point de vue disciplinaire choisi. Rabat [Rab] propose une synthèse des définitions existantes et écrit « *la première acceptation commune peut être qualifiée d'information "toute donnée porteuse de sens". L'information est comprise ici comme une virtualité (de sens et d'usage), existant en soi, identifiable comme une petite portion de réalité (isolable et discrète - comme disent les linguistes). C'est oublier que l'information est toujours saisie par une intelligence qui l'informe (l'inscrit dans une nouvelle forme) et lui donne son sens particulier - et cela quel que soit le degré d'expertise de celui qui perçoit. Selon ce point de vue l'information n'est pas initialement porteuse de sens ou plutôt est porteuse d'un sens virtuel qui demande à prendre corps au sein d'un contexte de lecture original et associé à une histoire du sujet* ». De nombreuses tentatives ont été faites pour opérer une synthèse entre la première acceptation (définition dite objective) et le caractère subjectif de la personne qui perçoit cette information. Selon Rabat, trois invariants se détachent de ces définitions :

- il existe bien un objet, en l'occurrence un état de connaissance informé, inscrit et destiné à être transmis,
- il convient d'envisager cet état de connaissance sous l'angle de la production du sens,
- l'ensemble est processuel et s'articule autour d'une situation de communication.

Il propose deux définitions :

- pour l'information décrite comme une propriété : élément de connaissance susceptible d'être représenté à l'aide de convention pour être conservé, traité ou communiqué.
- pour l'information conçue comme processus signifiant : l'information c'est le processus signifiant qui associe au sein d'un même message le producteur d'information et celui qui l'interprète. Il s'agit bien d'une « association » au sens fort, de type contractuel , entre une intention signifiante (producteur) et un projet de signification (lecteur/créateur de sens).

Dans ce travail de recherche, nous nous plaçons dans le cadre de la deuxième définition.

1.2 Les informations utiles à l'entreprise

Lesca et Lesca [LL10] distinguent trois types d'informations dans les entreprises :

- les informations de fonctionnement : elles sont indispensables au fonctionnement quotidien de l'entreprise, elles sont répétitives et formalisées,
- les informations d'influence : leur finalité est d'influer sur les acteurs pertinents de l'entreprise ; cette influence peut être exercée par des personnes internes ou externes à l'entreprise,
- les informations d'anticipation : elles permettent d'anticiper certains changements de l'environnement socio-économique ; ces informations sont peu répétitives, incertaines, ambiguës, fragmentaires, contradictoires ou encore mensongères. « *Une information anticipative permet de "voir venir à l'avance" un possible danger (ou une possible opportunité d'affaire)* »[LL11].

Dans ce travail, nous nous intéressons plus particulièrement au dernier type d'information. Lesca et Lesca [LL11] définissent deux types d'information anticipative :

– les informations de type « information de potentiel » : « *elles renseignent sur les potentiels et les faiblesses des acteurs situés dans l'environnement de l'entreprise et pertinents pour celle-ci. [...] Selon l'interprétation qui en est faite, l'information de potentiel nous renseigne sur les capacités à agir, d'un acteur pertinent de l'environnement* »[LL11],
– les informations de nature à déclencher une alerte : ce sont des informations dynamiques, furtives et difficiles à obtenir. Parmi elles, se trouvent les signaux faibles.

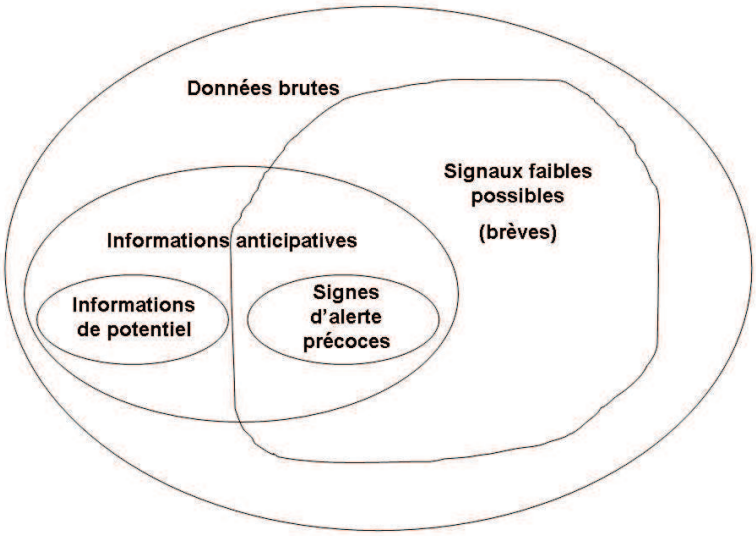


Figure I.1 – Classement des informations tiré de [LL11]

Dans le cadre de la veille stratégique, les signaux faibles ont une grande importance.

1.3 Signaux faibles

L'expression « weak signal » traduite par « signal faible » a été employée pour la première fois par Ansoff [Ans75]. Le terme « weak » ne signifie pas que le contenu du signal soit faible mais que sa forme est souvent incomplète et insaisissable. Ansoff définit le signal faible « *comme le point de départ d'une amplification à propos de laquelle seulement une information partielle est disponible au moment où la réponse doit être fournie et qui doit au besoin être complétée avant que des impacts sur l'entreprise ne commencent à se manifester* »(Ansoff cité par [KC05]). Il affirme que si l'organisation est attentive aux signaux faibles émis par son environnement, elle peut

anticiper les surprises et les ruptures stratégiques. La définition donnée par Lesca et Lesca [LL11] précise qu'un signal faible est une « *information se présentant* ex ante *comme une donnée d'apparence anodine mais dont l'interprétation que l'on en fait peut déclencher une alerte pouvant indiquer que pourrait survenir un évènement susceptible d'avoir des conséquences considérables (en termes d'opportunité ou de risque). Après interprétation, le signal est qualifié de signe d'alerte précoce* ».

Les caractéristiques des signaux faibles se classent en 2 groupes [LL11] : celui des caractéristiques utiles et celui des caractéristiques regrettables. Les caractéristiques utiles pour qu'un signal faible soit sélectionné sont :

- anticipatif,
- pertinent par rapport à un sujet de préoccupation d'un manager ou par rapport à la cible de veille.

Les caractéristiques regrettables des signaux faibles rendent ce type d'information très délicat à détecter, à fiabiliser et à interpréter [LL11]. Elles justifient le mot « faible » (voir tableau I.1).

« Faible » parce que	**Justification du mot « faible »**	**Différence avec une information de fonctionnement**
Fragmentaire (Fragmentary)	Nous sommes en situation d'information incomplète. Nous ne disposons que d'un fragment d'information à partir duquel on pourra se risquer à faire des inductions dans une démarche de type holistique, par exemple.	L'information de gestion courante est complète
Noyé dans un océan de données brutes (Raw data)	Disséminé dans une multitude de données inutiles (raw data), de données qui « font du bruit », le signal faible risque de passer inaperçu. Le plus grand nombre des personnes passe à côté de cette information.	Claire et distincte

Signification non évidente (Equivocal)	Le signal faible est peu parlant en soi. Sa signification est souvent peu évidente et ambiguë. Certains auteurs utilisent le vocable « équivoque » (equivocal).	Exprimée dans un langage « codifié » dans l'entreprise
Insolite / fortuit / inattendu (Unusual, Unfamiliar, Unexpected)	Le caractère insolite d'un signal faible rend plus difficile son repérage. On ne s'attend pas à son apparition.	Répétitive, familière
Sans utilité apparente, (Useless Unnecessary)	Sans lien évident/apparent avec une préoccupation en cours. Le même signal faible peut interpeller une première personne et paraître sans intérêt évident pour son entourage professionnel. Son utilité ne saute pas aux yeux, les conséquences de l'événement « signalé » ne s'imposent pas d'elles-mêmes. La « valeur de cette information » n'est pas évidente au premier abord	Indispensable, pour effectuer une tâche ou résoudre un problème (généralement répétitif)
Peu visible, difficile à discerner (Noticing)	Un signal faible passe facilement inaperçu : il est furtif, fugace. La détection d'un signal faible ne consiste pas seulement dans une recherche d'information, elle demande aussi un apprentissage et une capacité de discernement.	Réclamée par son utilisateur
Isolé, singulier (Singular Isolated)	On ne sait pas à quoi relier le signal faible, dans quelle « catégorie mentale » ou dans quel dossier le classer ; éloigné des préoccupations du moment.	Fait partie d'un dossier complet et ordonné
Fiabilité douteuse (Confidence)	On s'interroge immédiatement sur la source du signal faible et sur la confiance qu'on peut lui accorder.	Vérifiée (en principe)

Aléatoire, asynchrone (Unforeseeable)	Un signal faible n'apparaît pas lorsqu'on le désire. Par rapport à un capteur, un signal faible apparaît de façon aléatoire. La source, ou encore l'émetteur du signal, est totalement indépendante du capteur. En outre, deux (ou plusieurs) signaux faibles concernant un même sujet, ne parviennent pas nécessairement dans un ordre logique ou chronologique.	Elément de processus de travail dans lequel les tâches sont coordonnées
Subjectif (Subjective)	Un signal faible jugé intéressant par une personne, semble sans intérêt pour les autres.	Régie par des procédures objectives
Qualitatif	Qualitatif le plus souvent.	Quantitative la plupart du temps
Format	Formes les plus diverses : écriture, dessin, photo, odeur, sensation du toucher, bruit, goût, etc. Non-dit, silence, etc.	Ecriture, numérique, de plus en plus souvent
Devant être recherché dans un environnement sujet à des discontinuités	Il est très difficile de construire des algorithmes permettant de détecter automatiquement des signaux faibles.	

Tableau I.1 – Tableau des caractéristiques des signaux faibles [LL11]

La notion de signal faible étant difficile à appréhender, il nous est apparu intéressant au stade de notre développement de donner des exemples.

Exemples de signaux faibles :

– « le cas Christine Bénard »[LL11] : dans un journal économique a été publiée la phrase suivante « Madame Ch. Bénard vient d'être nommée directrice achats du groupe Valeo ». Cette phrase noyée dans un journal aurait pu attirer l'attention d'un fournisseur de Valeo. Le secteur des équimentiers automobiles et plus particulièrement la fonction achats sont traditionnellement des milieux

très « machistes » : il est par conséquent surprenant qu'une femme soit nommée à un poste de responsabilité (directeur général au niveau holding). Nommer une femme à un poste haut placé peut signifier qu'elle bénéficie d'une faveur extraordinaire (dans ce cas-là aucun intérêt pour les fournisseurs) ou qu'elle a de très bonnes qualités pour occuper son poste. Dans ce deuxième cas, un scénario possible est l'apparition d'une politique de « réduction des coûts » qui pourrait entraîner de profonds changements dans les relations avec les fournisseurs, voire conduire à la suppression de ces dernières. Pouvoir anticiper un tel scénario est évidemment d'une importance capitale pour les fournisseurs.

– un chef d'entreprise X dans le secteur de la chimie passait près de chez son concurrent Y quand il a senti une odeur qui l'a intrigué. En arrivant à son bureau, il a compris pourquoi cette odeur avait attiré son attention : il s'agissait de celle d'un « ingrédient » d'un de ses produits pour lequel Y n'était pas un concurrent. Par conséquent ce chef d'entreprise en a conclu que Y allait devenir également un concurrent sur ce produit.

Ces deux exemples montrent que lorsqu'un signal faible est perçu et collecté, il est nécessaire de l'interpréter et de lui donner du sens. Sans cette interprétation le signal n'a aucune valeur. Pour bien interpréter des informations, il est nécessaire de connaitre et d'analyser la source qui les produit.

1.4 Quelles sont les sources d'informations ?

Il existe différentes typologies des sources d'informations qui dépendent de l'origine, du type, du contenu, de la structure ou de la distance de l'auteur.
La première différencie l'origine des informations :

– **les sources externes** : les informations sont issues de l'environnement extérieur de l'entreprise (organisme public, fournisseur...),
– **les sources internes** : les informations proviennent de l'organisation elle-même et de ses différents services.

D'autre part, les sources peuvent être classées en fonction de la nature de l'information :

– les supports de l'information textuelle,
– les supports de l'information sonore,

- les supports de l'information visuelle,
- les supports de l'information multimédia.

Les sources peuvent être également distinguées par rapport à leur distance de l'auteur (sources primaires et secondaires) et par rapport à leur structure (sources formelles et informelles), d'après Medeiros Wanderley [MW04] :

- **Sources primaires** : l'information issue de ces sources est brute, elle n'a pas subi de transformation. Les points de vente (informations recueillies grâce aux caisses à lecture laser), les tests ou observations sont des exemples de sources primaires
- **Sources secondaires** : les informations ont été préalablement collectées, traitées et analysées, ce sont des informations de « seconde main » ; par exemple les revues de spécialistes ou les organismes publics (tel que l'INSEE) fournissent des informations analysées.
- **Sources d'information formelles** : « *On entend par informations formelles les publications scientifiques et techniques, les brevets, les livres. Ce sont des informations qui généralement sont validées. Cependant, leur durée de vie est variable du fait de l'évolution des contraintes et des normes, du fait aussi que les informations publiées sont toujours en retard sur la pensée du moment (ceci à cause du travail et des délais matériels préalables à la publication, ce laps de temps pouvant aller jusqu'à deux ans)* »[Dou05]. Les sources formelles sont par exemple : la presse (journaux, magazines, télévision, cinéma, radio), les publications scientifiques et techniques, les publications d'entreprise (rapport annuel, offres d'emploi, ...) les banques de données, les brevets, les tribunaux de commerce, cadastre, Internet...
- **Sources d'information informelles**
  Lorsqu'une information provient d'une source informelle, on parle d'« information de terrain ». Cette expression « *désigne une information dont la source est un homme en action sur le terrain. Il capte une information. [...] Cette information est généralement d'origine sensorielle : une observation visuelle, une phrase entendue, une odeur insolite, ...* »[LL11]. Parmi les sources informelles, on trouve les produits des concurrents, les fournisseurs, les sous-traitants, les clients, les missions et voyages d'étude, les expositions et les salons, les col-

loques et congrès, les contrats de recherche, les mémoires d'étudiants...

Afin d'obtenir des informations anticipatives, Lesca [Les03a] préconise d'accorder davantage d'attention aux informations issues de sources primaires et informelles sans écarter les sources formelles. Les informations de sources primaires et informelles, nommées information de terrain par Lesca [Les03a], sont non répétitives, ambiguës, incomplètes et fragmentaires [JMFL06] : par exemple le traqueur aura vu, entendu, senti ou même goûté « quelque chose » qui l'aura frappé.
Les signaux faibles, collectés auprès de sources diverses, sont utiles pour les dirigeants d'entreprise afin d'anticiper des opportunités ou des risques. Dans cette section, nous avons défini le concept de signal faible mais cela ne suffit pas pour les entreprises pour agir. Il faut des méthodes et des outils pour collecter et utiliser les signaux faibles.

2 Comment collecter et utiliser des informations et les traiter ?

2.1 Définitions de la veille stratégique

L'expression « veille stratégique » est une traduction de « environmental scanning ». Depuis les travaux d'Aguilar [Agu67], plusieurs auteurs se sont intéressés à cette activité cruciale pour l'entreprise afin de gérer l'incertitude créée par l'environnement. Le tableau I.2 tiré de Kamoun-Chouk[KC05] présente l'évolution dans le temps de la définition de ce concept : « *d'une simple activité d'acquisition de l'information environnementale la veille stratégique est passée à une activité plus complète intégrant l'interprétation des informations dans une perspective d'anticipation et l'insertion des résultats dans le processus de décision stratégique au sens large (et non pas strictement réservé au directeur de la stratégie)* »[KC05]. Ce tableau présente la veille stratégique comme :

- une démarche d'adaptation à l'environnement,
- un processus de collecte et d'interprétation des informations relatives aux évènements et tendances de l'environnement,
- une activité destinée à soutenir la décision stratégique,

Auteurs	Définitions
Aguilar, F.J. 1967[Agu67]	Scanning is the activity of acquiring information
Thompson, J.D. 1967 [Tho03] ; Pfeffer J. et Salancik G.A. 1978[PS78] ; Culnan, M. J. 1983 [Cul83]	Scanning, or the acquisition of information about events occuring outside the organisation, is one strategy that an organisation may employ in order to respond effectively to changes in the environment
Aguilar, F.J. 1967[Agu67] ; Choo C.W., Auster E. 1993 [CA93] ; Auster E., Choo C.W. 1994[AC94]	environmental scanning is defined as the acquisition and use of information about events and trends in an organisational external environment, the knowledge of which would assist management in planning the organization futur course of action
Aguilar, F.J. 1967 [Agu67] ; Sawyerr O.O. et Mc Gee J. E. 1999 [SM99]	Environmental scanning is the process of monitoring the external environment and collecting information of strategic importance for use in making organisational decision
Hambrick, D.C. 1981[Ham81]	The scanning is the first link in the chain of perception and actions that permit an organisation to adapt to its environment
Hambrick, D.C. 1982[Ham82] ; Daft R.L., Weick, K.E. 1984 [DW84] ; Farh J. et al. 1984 [FHH84] ; May R.C et al. 2000 [MSJS00]	Environmental scanning, the search mechanism by which managers discover important events and trends outside their organizations is the first step in this problem solving sequence
Hambrick, D.C. 1982[Ham82] ; Culnan, M.J. 1983 [Cul83] ; Daft et al. 1988 [DSP88] ; Elenkov, D.S. 1997 [Ele97]	Environmental scanning is the means through which top managers perceive external events and trends
Lenz R.T, Engledow J. L 1986 [LE86] ; Smeltzer L.R. et al. 1988. [SFN88]	Environmental scanning is defined as gathering and interpreting pertinent information and introducing the results into the organisational decision process
Ghoshal , S . 1988 [Gho88]	Environmental scanning is the activity by which organizations collect information about their environments

Tableau I.2 – Définitions successives de la veille tirée de Kamoun-Chouk[KC05]

– un premier pas dans le processus de résolution de problème.

Nous retenons, pour ce travail, la définition proposée par Lesca [Les03a] : la veille stratégique anticipative est définie comme un « *processus collectif et proactif par lequel des membres de l'entreprise traquent (perçoivent et choisissent) de façon volontariste, et utilisent des informations à caractère anticipatif et pertinentes concernant leur environnement extérieur et les changements pouvant s'y produire* ». Pour beaucoup la veille stratégique est synonyme d'intelligence économique or ce n'est pas le cas. Selon Lesca[Les03b], « *la veille stratégique s'en tient exclusivement à scruter et exploiter l'information accessible dans l'environnement extérieur de l'entreprise. L'intelligence économique intègre pleinement cette activité de veille stratégique, mais lui adjoint d'autres activités spécifiques [...]. La veille est donc un sous-processus de l'intelligence économique* ».

2.2 Comment faire de la veille ?

D'après Guechtouli [Gue09], la plupart des modèles de veille stratégique proposés dans la littérature repose sur la démarche du « cycle du renseignement »(voir figure I.2). Selon cet auteur [Gue09], cette démarche comprend les étapes suivantes [SA97] :

– l'expression des besoins ;
– la recherche d'informations ;
– l'exploitation (vérification, traitement, analyse, synthèse) ;
– la diffusion des destinataires.

Ce schéma, très utilisé, reste cependant incomplet et réducteur. Afin de rendre ce cycle du renseignement opérationnel dans le cadre de la veille, plusieurs processus ont été proposés. La figure I.3 présente le processus décrit par la norme de l'AFNOR [AFN98]. Selon Favier et Ihadjadene [FI04], « *cette norme ne répond pas directement à la question de savoir ce que doit être un système du point de vue de l'usager ayant tel besoin d'information mais ce qu'il doit être dans le cadre d'un contrat liant un prestataire à un client du point de vue du prestataire* ». Kengen [KEN05] met en avant que dans le processus modélisé par l'AFNOR (figure I.3) :

– le prestataire n'est pas tenu de divulguer ses sources au client dès lors que les « axes de surveillance » auront été définis.
– le prestataire doit dégager du sens des informations collectées et proposer une

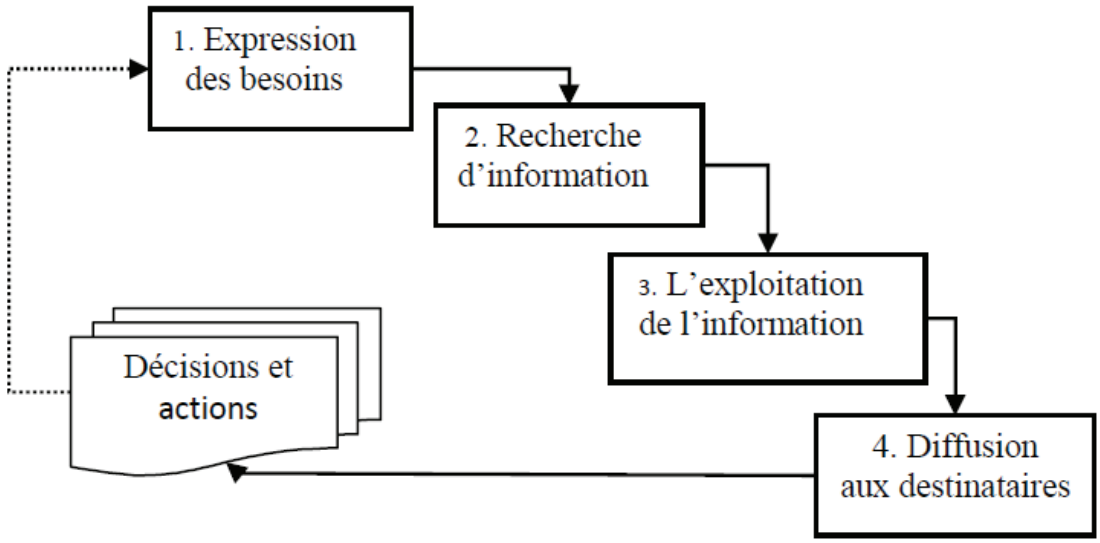


Figure I.2 – Etapes du processus de veille selon le cycle du renseignement [Gue09]

formulation adaptée au processus de décision du « client ».

- « *la communication des résultats peut être l'occasion d'un ajustement par approfondissement et/ou réorientation des objectifs et moyens de veille.* » Rien n'est donc figé et la veille est reconnue comme un processus évolutif.
- est mise en avant une volonté claire, dans le chef des concepteurs de la norme, d'encourager la délivrance d'un produit à valeur ajoutée.

Contrairement à cette norme AFNOR, Lesca [Les03a] propose un modèle, appelé VASIC (veille anticipative et intelligence collective), qui s'intègre dans l'organisation sans faire appel à un prestataire.

2.3 Le processus VASIC

Le processus proposé par l'équipe de veille stratégique du CERAG est illustré par la figure I.4.

2.3.1 Ciblage

Le ciblage a pour objectif la délimitation d'une « *partie de l'environnement sur laquelle les dirigeants de l'entreprise ont jugé pertinent de porter leur attention en*

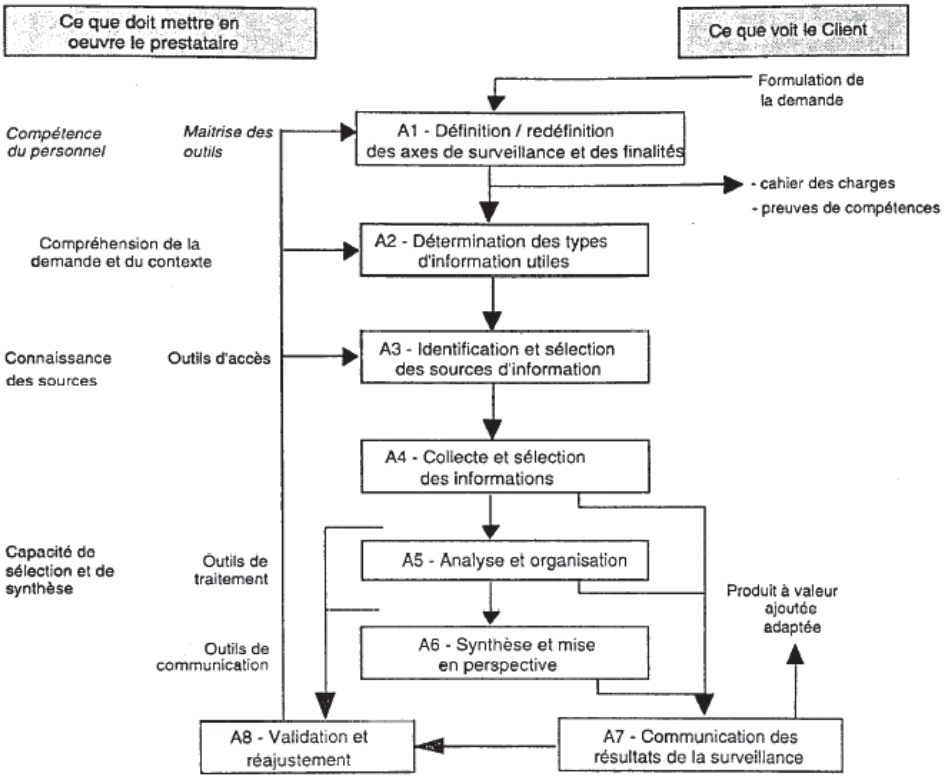


Figure I.3 – Modèle de veille stratégique selon l'AFNOR[AFN98]

priorité et pour une période donnée »[LL11]. La cible est construite en six étapes [Les03a] :

1. délimiter les domaines d'activité de l'entreprise à placer sous surveillance,
2. lister les acteurs [2] extérieurs pertinents à placer sous surveillance et lister les thèmes [3] pertinents pour chacun des acteurs choisis,
3. consulter en interne pour valider les listes d'acteurs et de thèmes,
4. hiérarchiser les acteurs listés, sélectionner des acteurs prioritaires et hiérarchiser les thèmes prioritaires pour chacun des acteurs ciblés,
5. cibler les informations à traiter pour chacun des thèmes,
6. cibler les sources d'informations.

2. un acteur est un individu ou groupe d'individus extérieurs à l'entreprise et dont les décisions et les actions sont susceptibles d'influencer le devenir de l'entreprise
3. centre d'intérêts pour l'entreprise concernant un ou plusieurs acteurs

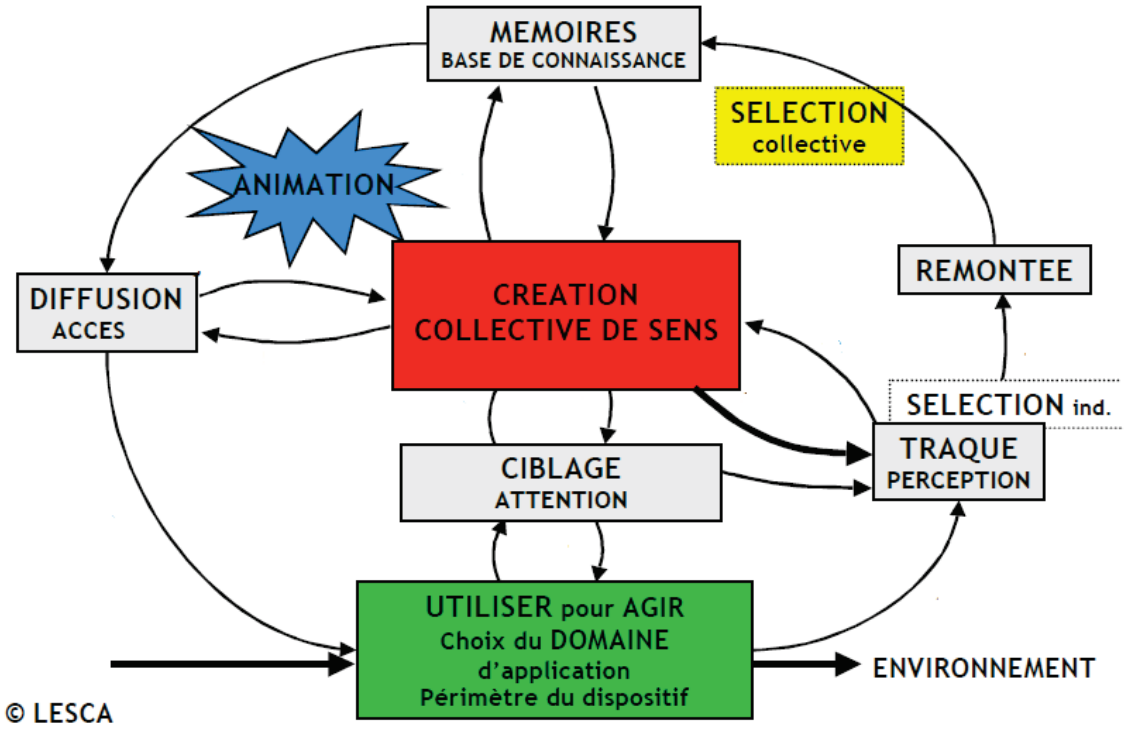


Figure I.4 – Modèle générique adaptatif VASIC

Une fois la cible déterminée, il est possible de capter des informations à son sujet.

2.3.2 Captage/perception

On parle aussi de « traque » pour désigner la captage d'informations dans l'entreprise. « *La traque est l'opération volontariste (proactive) par laquelle des membres de notre entreprise perçoivent, choisissent (élisent) ou provoquent des informations. Nous avons choisi le mot "traque" pour signifier que les informations de VASIC les plus intéressantes ne viennent pas d'elles-mêmes à nous* »[Les03a].

Cette phase de captage est réalisée par des personnes appelées « traqueurs ». Ils sont de deux types :

- les traqueurs dits « itinérants » : ils sont en contact direct avec l'environnement extérieur (commerciaux, acheteurs, techniciens...) grâce à leurs nombreux déplacements.
- les traqueurs dits « sédentaires » : leur principale mission est de traquer des

informations en accédant à diverses bases de données, des journaux, Internet... Ce sont des professionnels de l'information.

Les traqueurs vont remonter l'information vers les mémoires et la mise en commun des informations. Les informations traquées peuvent être en grand nombre et doivent par conséquent faire l'objet d'une sélection collective.

2.3.3 Sélection collective des informations

« *La sélection des informations est l'opération qui consiste à ne retenir, parmi les informations recueillies ou se présentant à nous, que les seules informations susceptibles d'intéresser les utilisateurs potentiels. Cette opération est cruciale : une absence de sélection conduit à "trop d'informations" (information overload [EM04]) et à étouffer le processus VASIC, tandis qu'une sélection trop restrictive appauvrit et assèche le processus VASIC* »[Les03a]. Un groupe de travail est constitué pour examiner les informations remontées par les traqueurs. Pour réaliser leur sélection, les participants s'appuient sur leurs connaissances tacites, les connaissances formelles de l'entreprise et les différentes mémoires de l'entreprise.

2.3.4 Mémoires

Le mot « mémoires » désigne « *tous les lieux, quels qu'ils soient, où peut se situer une information ou une connaissance. Dans notre contexte de l'Intelligence Collective d'entreprise, nous voulons désigner toutes les mémoires, quelles que soient leurs formes, concernant le dispositif VASIC : il s'agit de la mémoire des personnes (ce que chacune d'elles a en tête), des dossiers formels, des mémoires informatisées,...* »[Les03a]. L'entreprise peut, par exemple, mettre en place des outils permettant aux traqueurs de sauvegarder leurs informations, par exemple s'il s'agit d'articles de presse. Du fait de la diversité des mémoires, il est nécessaire de mettre en place une intelligence collective permettant le rapprochement et l'interprétation des informations issues de ces différentes mémoires.

2.3.5 Création collective de sens

« *La "création collective de sens" est l'opération collective au cours de laquelle sont créés de la connaissance et du "sens ajouté", à partir de certaines informations*

qui jouent le rôle de stimuli inducteurs, et au moyen d'interactions entre les participants à la séance de travail collectif, ainsi que entre les participants et les diverses mémoires (tacites et formelles) de l'entreprise. Le résultat de la création collective de sens est la formulation de conclusions plausibles (hypothèses) devant déboucher sur des actions concrètes »[Les03a]. Le déroulement d'une séance de création collective de sens sera détaillé à la section 3. Les conclusions des séances de création collective de sens doivent être diffusées aux personnes concernées.

2.3.6 Diffusion de l'information

« *C'est l'ensemble des opérations grâce auxquelles une information nouvelle (un commentaire, une connaissance utile pour l'action, etc.) parvient aux personnes (un utilisateur potentiel) qui sont censées les utiliser, y compris les traqueurs pour la part qui les concerne et qui les motive* »[Les03a]. Il s'agit de fournir la bonne information au bon moment à la bonne personne, la personne qui peut transformer l'information en action.

2.3.7 Animation

Selon Lesca [Les03a], « *animer consiste, ici, à "donner une âme", à insuffler la vie au processus VASIC, dont le moteur est essentiellement humain, et au dispositif organisationnel et technique qui en est le support. La personne chargée de cette animation est appelée Animateur* ». Il est essentiel que le processus de VASIC soit animé par un responsable officiel, clairement désigné et reconnu par la hiérarchie ; sinon le dispositif a de fortes chances de ne pas fonctionner [LL11].

Le cœur de la thèse porte sur les outils permettant la mise en place de la séance de création collective de sens et notamment la phase préparatoire qui fait partie des missions de l'animateur[ML10].

3 Création collective de sens

3.1 Déroulement

Pour aider les entreprises à créer collectivement du sens, l'équipe de veille stratégique du CERAG a proposé une méthode nommée « puzzle »[Les92]. L'idée à l'origine est la métaphore du jeu de puzzle. Lors de la séance de création collective de sens, l'animateur apporte 5 ou 6 brèves (« *une information ramenée à ses mots essentiels de façon à être très courte. Cette contrainte de taille résulte du fait qu'une brève est destinée à être projetée sur un écran* »[LL11]). Ces brèves vont être utilisées comme des pièces d'un puzzle à construire. Contrairement au jeu de puzzle, on ne dispose pas d'un modèle ni de toutes les pièces.« *La méthode puzzle consiste à construire collectivement un puzzle (ou plusieurs si nécessaire), sur l'écran de la salle de travail, en utilisant des signaux faibles en guise de pièces, d'une part, et les suggestions et connaissances tacites que les participants explicitent alors, d'autre part* »[LL11]. Les participants proposent un placement sur l'écran du signal faible et le type de lien que ce signal faible peut avoir avec les autres.

Les figures I.5 et I.6 présentent un exemple de construction progressive d'un puzzle. Ce puzzle a été réalisé dans le cadre d'un projet avec le ministère de l'économie belge. Pour une séance de formation, 8 FULL texts (données brutes sous forme de textes) sur le domaine de l'agroalimentaire avaient été sélectionnés (les FULL texts se trouvent en annexe). Dans un premier temps, les participants ont lu ces FULL texts et en ont extrait 9 brèves. L'animateur a demandé par quelle brève on pouvait commencer le puzzle. Une fois la brève choisie, une seconde a été choisie. L'animateur a incité les participants à proposer un positionnement sur l'écran de la seconde brève par rapport à la première ainsi que le lien qu'il pouvait exister entre les deux. Toutes les brèves ont été étudiées de la même manière : placement sur l'écran et le lien avec les brèves déjà présentes. Les brèves nommées B (et sur fond blanc dans les figures I.5 et I.6) sont des brèves tirées des FULL texts et celles nommées C (fond gris) sont le résultat de fusion de brèves B. Les flèches rouges indiquent que les participants estiment que les deux brèves sont en contradiction.

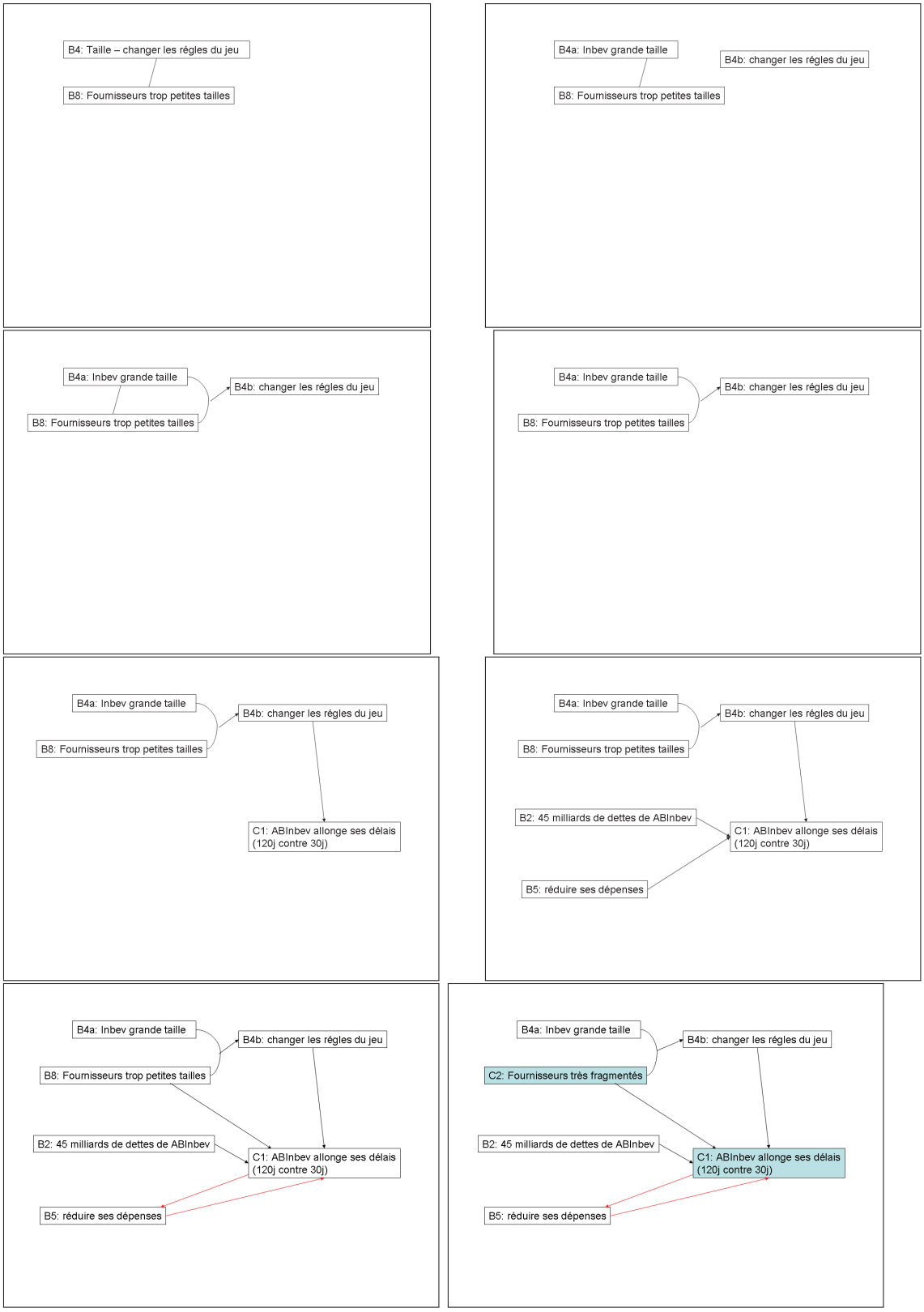


Figure I.5 – Évolution d'un puzzle au cours d'une séance de création de sens

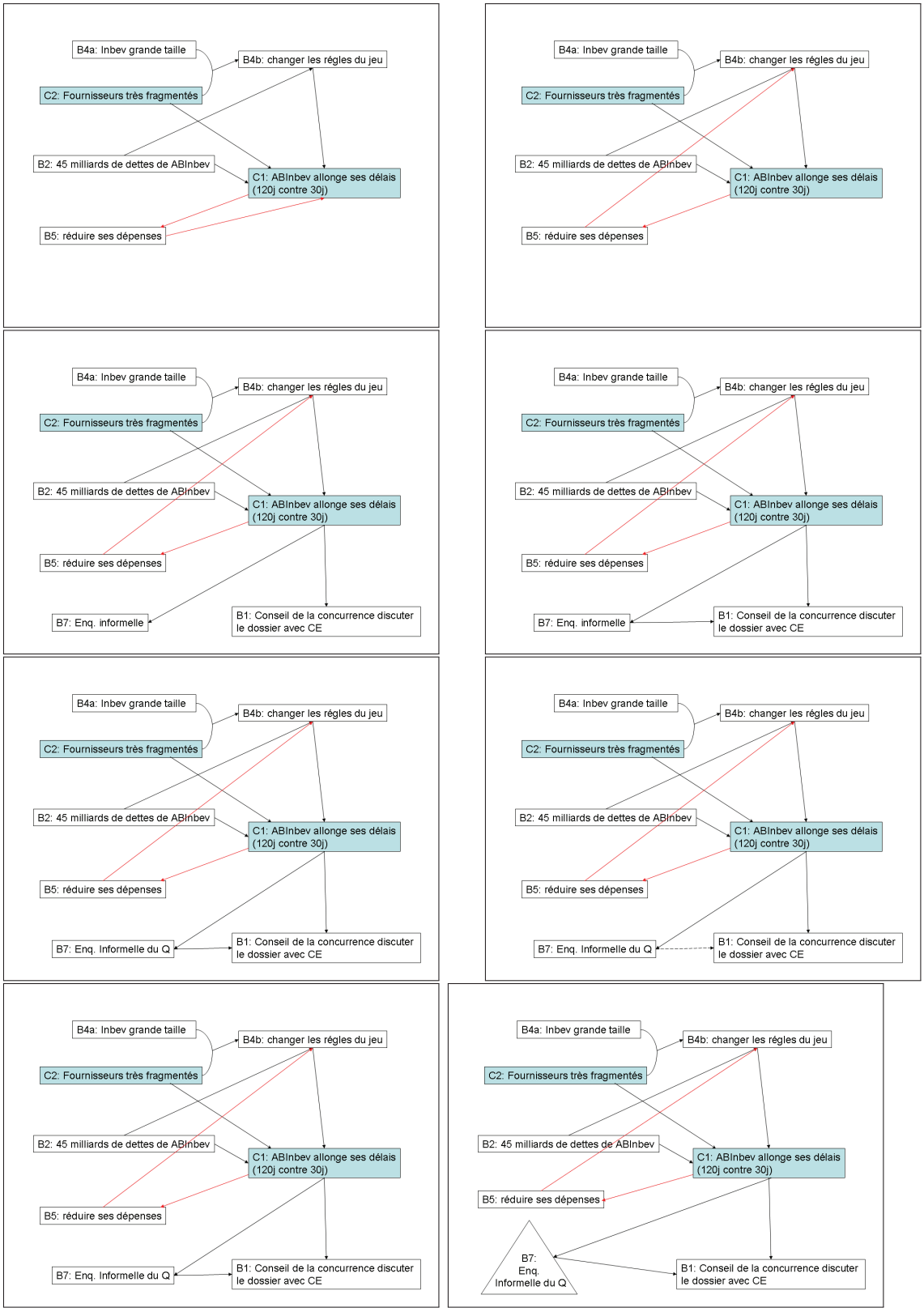


Figure I.6 – Évolution d'un puzzle au cours d'une séance de création de sens (suite)

Les apports de la méthode « Puzzle » sont [Les03a] :

- ouvrir l'imagination des participants en prenant pour point de départ des informations « disponibles », formelles et tacites, celles-ci jouant le rôle de stimuli,
- « voir entre les informations » disponibles et combler « les vides »,
- stimuler les interactions entre les participants,
- expliciter le raisonnement collectif ainsi effectué,
- fournir des aides à la visualisation des raisonnements,
- rendre le raisonnement communicable à d'autres personnes, hors du groupe de travail,
- garantir la traçabilité des raisonnements effectués collectivement,
- aider à mémoriser les raisonnements et les justificatifs des choix effectués collectivement.

Le résultat d'un séance de création collective est composé d'un ou plusieurs puzzles. Un « *puzzle désigne une construction graphique argumentée, constituée d'un petit nombre de fragments d'information se rapportant à une même problématique (acteur et/ou thème) et supposées s'enrichir mutuellement* »[LL11]. Les fragments d'information sont issus de FULL texts (ou « informations voisines ») sélectionnés par l'animateur lors de la préparation de la séance collective de sens.

Lors de la séance de création collective de sens, l'intelligence collective (intelligence d'un groupe d'individus) a un rôle central afin de créer de la connaissance pour agir.

3.2 Intelligence collective

Daft et Weick [DW84] sont à l'origine du concept d'« intelligence organisationelle ou collective ». Ils définissent les organisations comme des systèmes sociaux d'interprétation aboutissant à de nouvelles connaissances. Nonaka et Toyama [NT03] estiment que les organisations ne sont pas seulement une machine à traitement de l'information, mais une entité qui crée de la connaissance à travers l'action et l'interaction. Dans la méthode proposée par Lesca [Les03a], l'intelligence collective « *est appliqué[e] à un groupe de personnes qui acceptent volontairement de mettre en commun (en collectif) leurs capacités de détecter des évènements, d'en parler,*

de les interpréter et d'en tirer des enseignements utiles à l'action ». Garvin cité par Kamoun-Chouk [KC05] définit le concept d'organisation intelligente comme « *des endroits où les gens élargissent continuellement leur aptitude à créer les résultats qu'ils désirent réellement, des endroits où des nouveaux modes de pensée sont cultivés où l'aspiration collective est laissée libre d'agir à sa guise et où les gens ne cessent d'apprendre comment apprendre* ». Nonaka utilise le concept de Ba pour parler des « knowledge-creating place ».« *Ba can be thought of as a shared space for emerging relationships. This space can be physical (e.g., office, dispersed business space), virtual (e.g., e-mail, teleconference), mental (e.g., shared experiences, ideas, ideals), or any combination of them. [...] ba can be thought of as being built from a foundation knowledge* »[NK98](voir figure I.7). L'intelligence collective va permettre la création

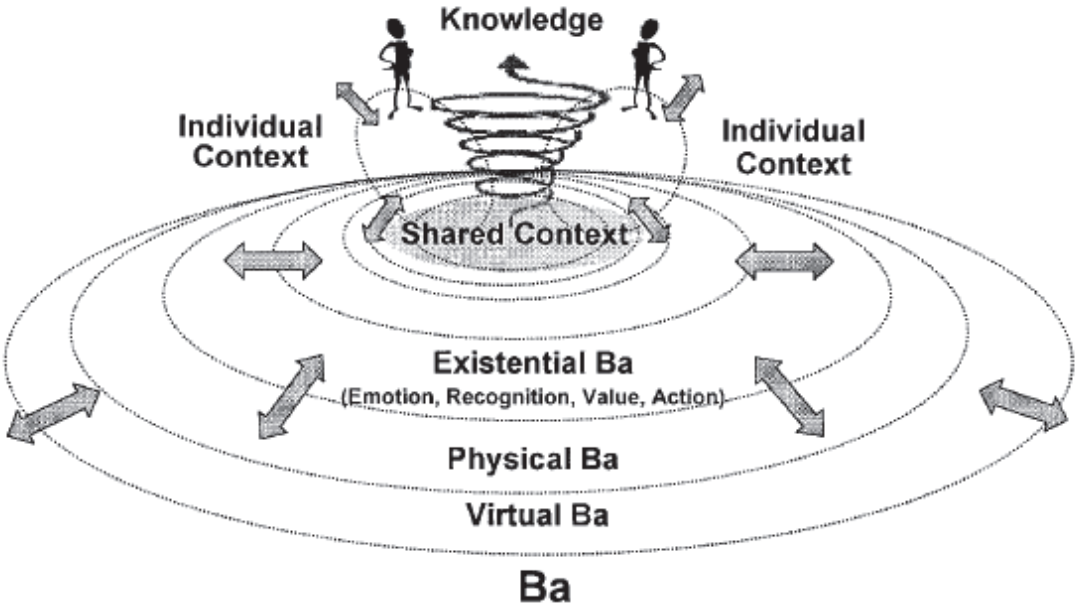


Figure I.7 – Représentation conceptuelle du Ba [NT03]

de sens. Weick[Wei79] et Huff[Huf90], cités par Caron-Fasan et Farastier [CFF03] soulignent que la construction ou création de sens est une étape essentielle dans la création de connaissances.

3.3 Création de connaissances

Le processus de veille stratégique permet de collecter de l'information qui vise à améliorer les connaissances de l'entreprise sur son environnement. Lesca et Dourai [LD10] définissent la « création de connaissance » comme « *l'opération collective par laquelle les informations sont "traitées" pour faire émerger du sens qui devra être utile pour l'action des responsables de l'entreprise* ». Selon Caron-Fasan et Farastier [CFF03], l'activité de veille stratégique s'inscrit dans un processus *continu, dynamique* et *évolutif* au sein duquel de nouvelles informations sont en permanence nécessaires. Le processus de veille est un processus d'apprentissage collectif et de création de connaissances organisationnelles.
Nonaka et Takeuchi [NT95] décrivent les modalités, selon lesquelles l'entreprise peut créer de la connaissance partagée à partir des connaissances individuelles de ses acteurs, à travers le concept de « spirale de la connaissance » qui traduit le passage des connaissances du niveau individuel au niveau collectif. Ils présentent un modèle se déroulant en spirale, appelé modèle SECI [NT95]. Il repose sur deux types de connaissances :

- les connaissances tacites : connaissances initialement non exprimées et difficilement exprimables (« nous en savons plus que nous ne pouvons en exprimer »). « *la connaissance tacite se construit en incorporant dans nos représentations mentales, les schémes d'actions qui ont obtenus des résultats efficaces (au sens où ils nous ont permis d'atteindre les objectifs poursuivis) dans la réalisation de tâches qu'elles soient cognitives ou concrètes. La connaissance tacite est donc fortement dépendante de notre perception du contexte de l'action, des caractéristiques de la tâche,des objectifs poursuivis, mais également de nos valeurs et non croyances. Elle est fondamentalement subjective, structurée et re-configurée en permanence par notre capacité à mettre en relation les informations acquises dans l'expérience personnelle, capacité fortement influencée par nos émotions et notre créativité* »[FB03],
- les connaissances explicites : elles sont formelles et systématiques. Pour cette raison, elles peuvent être facilement communiquées et partagées, par exemple les spécifications d'un produit ou une formule scientifique ou un programme informatique [Non91].

La figure I.8 présente le modèle de création de connaissances de Nonaka et Takeuchi.

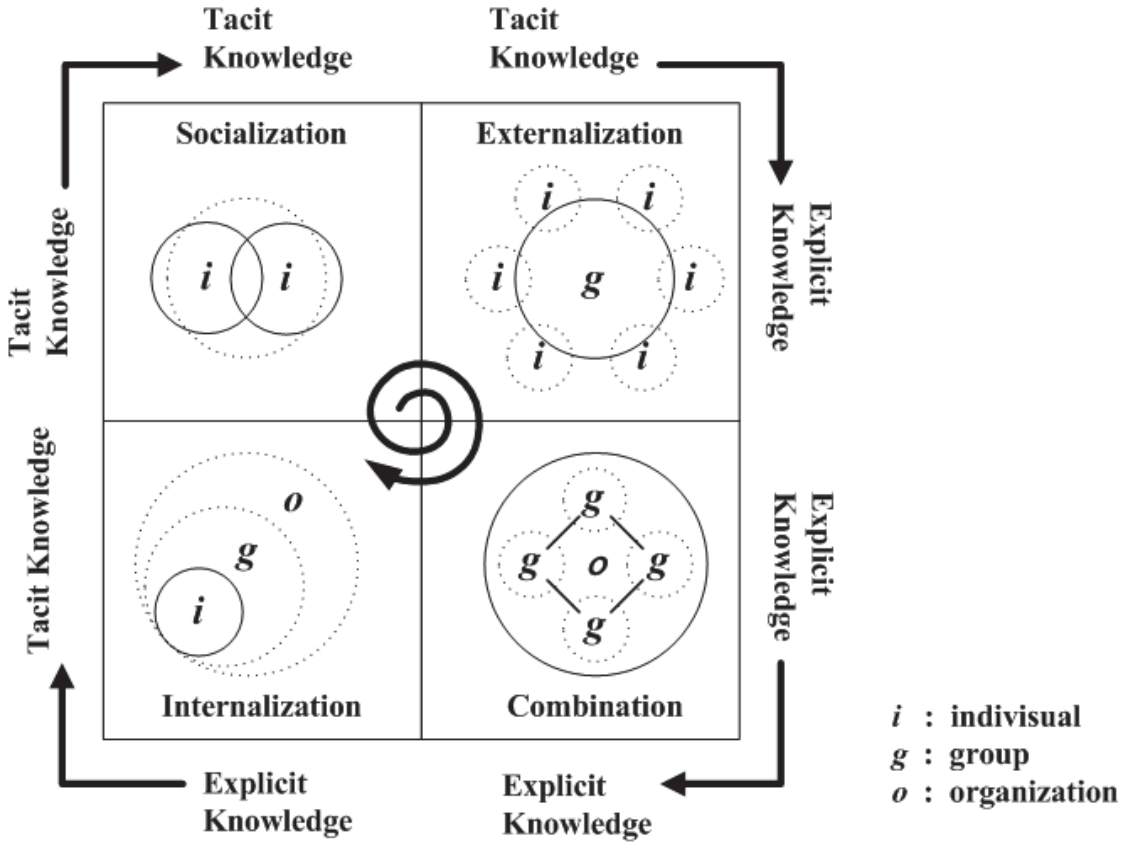


Figure I.8 – Modèle SECI [NK98]

La formalisation du modèle SECI est un processus en spirale d'interactions entre connaissances explicites et connaissances tacites. Il est composé des 4 phases :
Socialisation : dans cette phase, les individus échangent leurs connaissances tacites. « *Tacit knowledge is exchanged through joint activities - such as being together, spending time, living in the same environment - rather than through written or verbal instructions* »[NK98]. « *La connaissance tacite d'une personne ou d'un groupe peut alors devenir la connaissance d'autres personnes par des phénomènes tels que l'observation et l'imitation* » [CFF03].
Externalisation : lors de cette phase, les connaissances tacites vont être transformées en connaissances explicites à l'aide de métaphores, analogies, concepts, hy-

pothèses ou modèles. « *individuals use their discursive consciousness and try to rationalize and articulate the world that surrounds them. Here, dialogue is an effective method to articulate one's tacit knowledge and share the articulated knowledge with others. Through dialogues among individuals, contradictions between one's tacit knowledge and the structure, or contradictions among tacit knowledge of individuals are made explicit and synthesized* »[NT03].

Combinaison : cette phase implique la transformation de connaissances explicites en des ensembles plus complexes de connaissances explicites. Ces nouvelles connaissances sont disséminées dans l'organisation par des présentations ou des réunions.

Internalisation : « *the internalization of newly created knowledge is the conversion of explicit knowledge into the organization's tacit knowledge* »[NK98]. « *Par répétition, on enracine la connaissance explicite dans des séquences pouvant atteindre le stade du réflexe en adaptant le schéma explicite aux conditions spécifiques de l'exécution. L'internalisation de connaissances explicites peut aussi se produire par l'écoute ou la lecture de descriptions d'expériences passées* »[CFF03].

3.4 Séance de création collective de sens et création de connaissance

Caron-Fasan et Farastier [CFF03] ont montré que « *la veille stratégique, telle que décrite dans les travaux de Lesca, est un processus organisationnel ayant pour finalité la création de sens sur l'environnement de l'entreprise. Ce processus est décomposable en différentes phases qui, chacune, manipule, crée, transforme des connaissances tacites et explicites* » (voir figure I.9). La phase de ciblage réunit les décideurs qui en s'appuyant sur leurs connaissances de l'entreprise et de l'environnement vont définir leur cible. Cette étape peut être considérée comme une phase d'externalisation. Les étapes de captage et de sélection de l'information sont des phases qui nécessitent des connaissances tacites et explicites donc elles couvrent les phases de socialisation et d'externalisation. L'étape de diffusion de l'information s'inscrit dans l'internalisation de la connaissance.

Lors de la séance de création collective de sens :

- les participants vont échanger des connaissances tacites (socialisation),
- ils vont échanger de manière explicite leurs connaissances (externalisation),

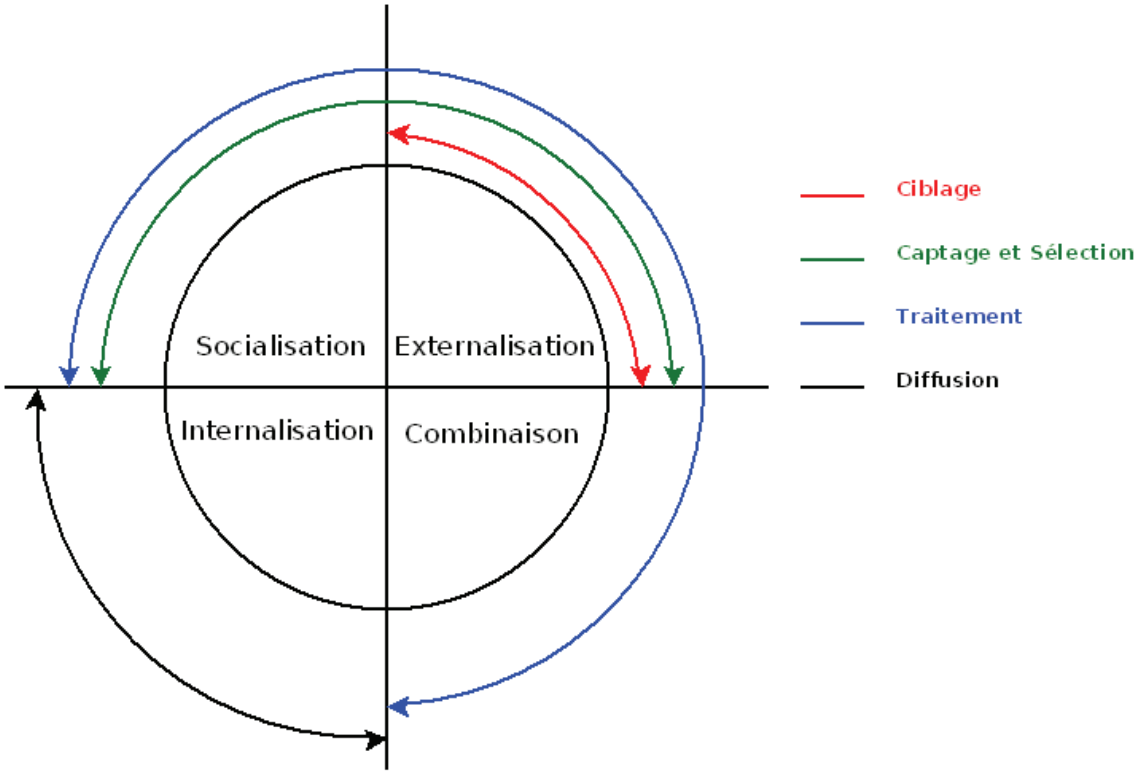


Figure I.9 – Transformation des connaissances durant le processus de veille stratégique d'après [CFF03]

– ils vont combiner les connaissances explicites issues de la sélection des informations (combination).

Pour mettre en place une intelligence collective afin de créer de la connaissance, l'animateur doit avoir sélectionné des informations lors de la préparation de la séance de création collective de sens.

4 Sélection des informations pour la séance de création collective de sens

Les traqueurs ont collecté de l'information concernant la cible choisie et cette information, sous forme de texte, est disponible dans une base. L'animateur dont c'est l'une des missions [ML10] doit préparer la séance de création collective de sens en sélectionnant des informations.

4.1 Le processus actuel de sélection

4.1.1 Processus

La sélection des informations (pour la séance de création collective de sens), illustrée par la figure I.10, est un processus itératif basé sur 4 étapes [Les10] :

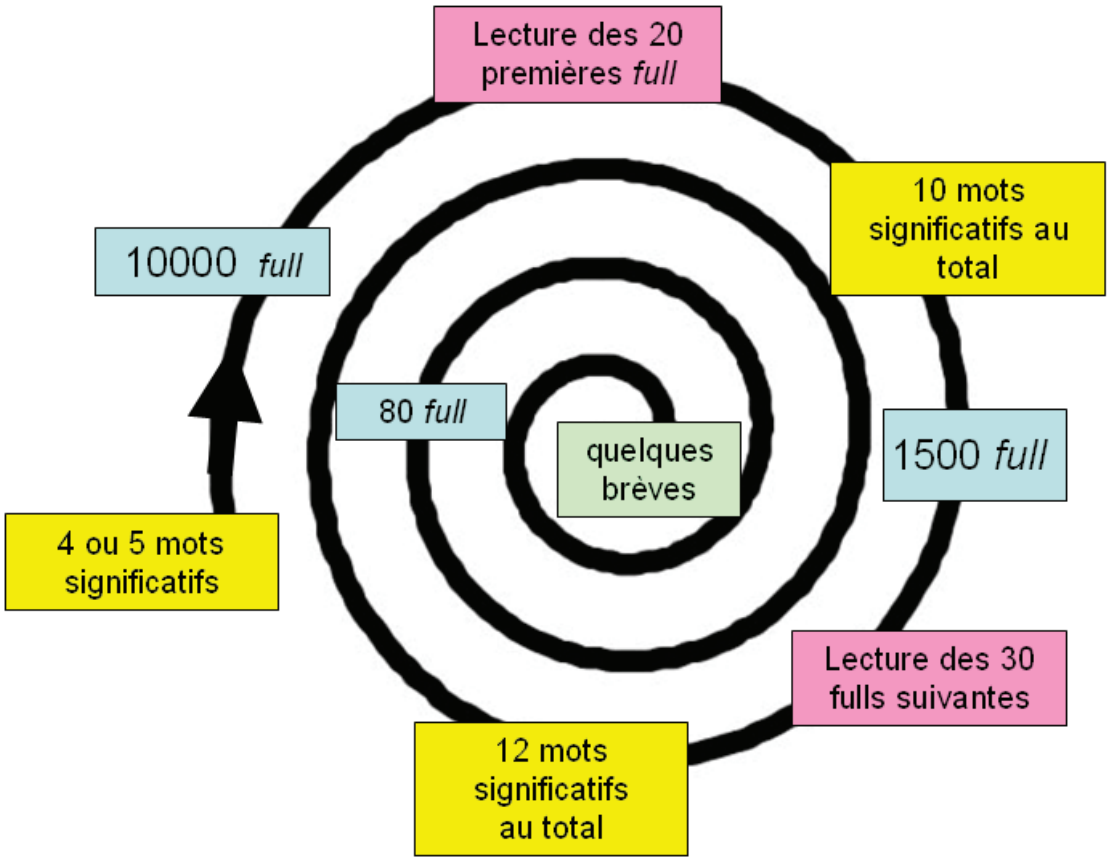


Figure I.10 – Spirale convergente des résultats [LL11]

1. recherche de FULL texts : à l'aide de 4 ou 5 mots significatifs qui définissent son besoin, l'animateur va rechercher dans la base des FULL texts (constituée par les traqueurs) ou sur Internet des informations. Il va obtenir un grand nombre d'informations qui ne pourront pas être utilisées lors de la séance de création collective de sens et il va donc devoir « distiller » l'information,
2. distillation de l'information (voir figure I.11) : chaque FULL text est analysé et une ou plusieurs brèves sont extraites ainsi que des mots significatifs

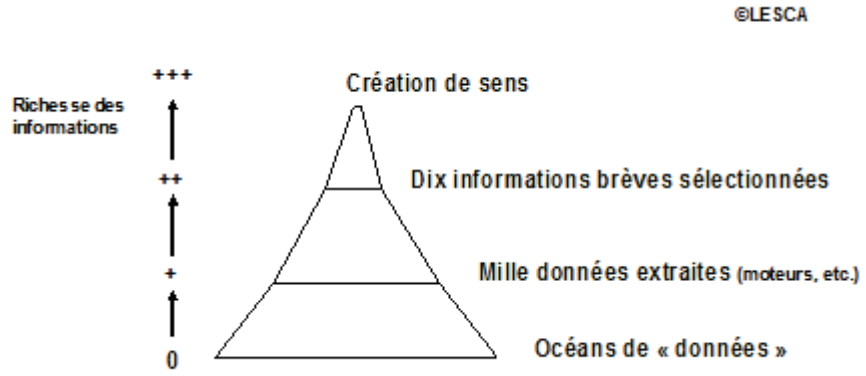


Figure I.11 – Distillation de l'information

3. constitution d'une liste de mots significatifs : chaque mot significatif extrait d'un FULL text est ajouté à une liste. Les caractéristiques de cette liste sont [Les10] :
   - le choix des mots est subjectif : il est propre à la personne qui fait le choix (par différence avec un thésaurus « officiel » de mots-clés),
   - la liste est à usage personnel,
   - le nombre de mots de la liste, pour un domaine donné, varie par ajouts ou suppressions de mots, au fur et à mesure qu'augmente l'apprentissage de la personne qui fait l'analyse,
   - cette liste ne sera jamais complète,
   - la personne qui établit la liste doit pouvoir justifier le choix de chacun des mots.
4. détection possible des signaux faibles : la distillation a pour objectif idéal l'identification de signaux faibles.

4.1.2 Difficultés et limites

La méthode Puzzle, appliquée des dizaines de fois dans divers organismes, a toujours donné satisfaction. En revanche, ces mêmes interlocuteurs ont clairement indiqué que cette méthode ne pourrait pas être durablement utilisée dans leurs organismes : la préparation des informations, dites informations brèves, demande trop de temps de travail humain. Alors que la recherche d'informations FULL texts est désormais très rapide sur l'Internet, la préparation des brèves (c'est-à-dire le filtrage

des données FULL texts pertinentes, la sélection des données en relation avec le sujet à traiter, l'identification du passage bref utile pour produire une brève elle-même utile pour la création collective de sens), ne peut être faite que manuellement, du moins pour le moment, et nécessite trop de travail et donc des coûts importants.

Une recherche à l'aide des mots significatifs sur des moteurs de recherche tels que Google, Yahoo ou Bing nous donne un très grand nombre d'informations. On parle alors de « surcharge d'information »[EM04]. Dans [LKC09], « *l'Internet se révèle comme un facteur d'échec potentiel de la veille du fait des maillons informatiques qui manquent dans le processus en aval de l'utilisation des moteurs de recherche* ». Les maillons manquants sont :

- incapacité, pour le moment, d'extraire des brèves en partant des données numériques volumineuses fournies par Internet,
- incapacité, pour le moment, de générer automatiquement des liens sémantiques entre les brèves alors que les utilisateurs croient que l'informatique permet désormais une « veille presse bouton »,
- impérialisme de l'Internet et relégation des informations d'origine terrain. « *La facilité avec laquelle l'Internet permet d'acquérir de gros volumes de données numériques (directement exploitables sur ordinateur) auprès de sources informatiques, a pour résultat de détourner complètement l'attention des informations d'origine "homme de terrain" qui sont pourtant potentiellement riches en signaux d'alerte précoce* »[LKC09].

4.2 Problématique de la veille stratégique

La problématique se situe à deux moments successifs du processus visant à détecter puis à interpréter des signaux faibles pour créer du sens. Le premier moment est celui où l'animateur doit extraire les données brutes FULL texts d'une vaste base de données. Ces FULL texts seront utilisés au cours d'une réunion de travail collectif dont les participants sont des membres de la hiérarchie de l'entreprise. Ceux-ci doivent faire le point sur une question appelée « ordre du jour » concernant la stratégie future de l'entreprise. Le second moment est la séance de travail collectif elle-même. Les participants sont susceptibles de poser diverses questions, concernant

les données utilisées, auxquelles l'animateur doit être en mesure de répondre sur le champ.

Pour préparer la réunion, l'animateur doit effectuer les tâches suivantes : rechercher et extraire les FULL texts se rapprochant de l'ordre du jour et susceptibles de contenir des signaux faibles annonciateurs de changements dans l'environnement de l'entreprise, répondre au sujet du degré de fiabilité de chacun d'eux, faire de nombreux allers et retours d'un FULL text à un autre pour accompagner les participants dans leurs réflexions et interactions.

Les dirigeants apprécient les séances de création collective de sens mais souhaitent que le temps de préparation et de manipulations des FULL texts soit le plus bref possible pour réduire les coûts administratifs d'une part, et ne pas gêner les réflexions d'autre part : sans un gain de temps significatif la veille stratégique sera abandonnée [Hem09][Les08] ! Ainsi la problématique se ramène à la question suivante : existe-t-il un logiciel capable de rechercher, dans une surcharge de données, les FULL texts pertinents et susceptibles de contenir de possibles signaux faibles, dans une base de données ?

Hypothèse de recherche : Si un logiciel permettait de rassembler rapidement/facilement les informations voisines d'une information donnée alors la veille anticipative stratégique utilisant les signaux faibles gagnerait en efficience (en termes de coûts et de satisfaction des utilisateurs).

4.3 Logiciels de veille existants

Afin d'apporter une solution satisfaisante pour sélectionner les informations, nous avons étudié les logiciels qui existent dans le domaine de la veille. De nombreux logiciels de veille ont été développés et la figure I.12 montre la variété des outils disponibles pour les veilleurs.
Selon François [Fra05], ces outils peuvent proposer jusqu'à 7 types de services :

- la collecte et la gestion de l'information :
  - les moteurs de recherche (Yahoo, Google, Exalead, ...)
  - outils de surveillance de source (AMI Market Intelligence, Website Watcher,

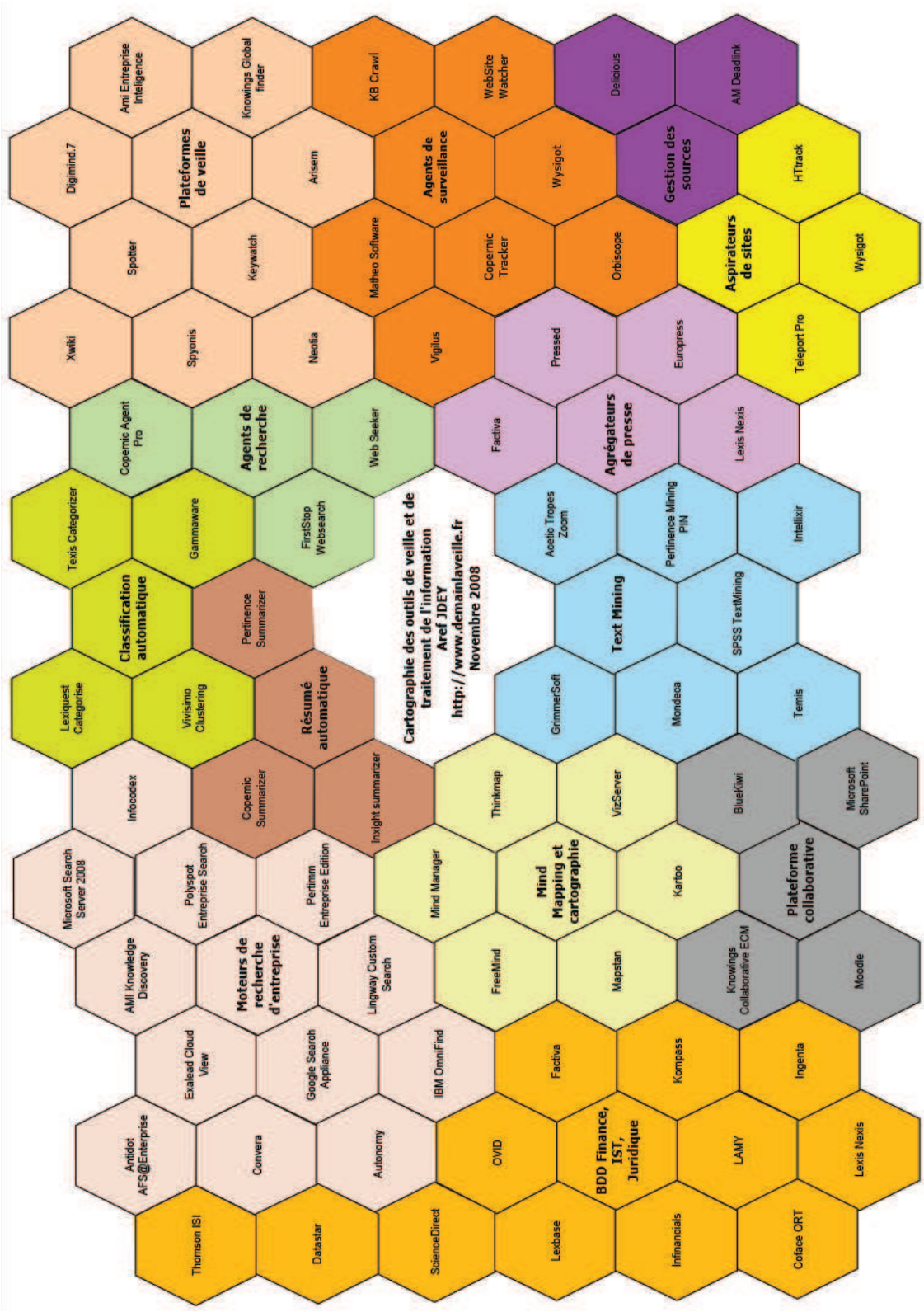


Figure I.12 – Cartographie des outils de veille et de traitement de l'information

...)
  - outils de gestion documentaire ou gestion des connaissances (Digimind Evolution, Keywatch, ...)
- l'analyse textuelle de l'information : « *l'analyse textuelle permet une indexation des textes qui sera ensuite utilisée pour catégoriser, classer ou cartographier des documents collectés* »[Fra05]. Les analyses textuelles sont réalisées selon deux approches (soit l'une, soit l'autre, soit en combinant les deux) :
  - approche statistique (Alceste, Tétralogie, Sphinx Lexica ...)
  - approche linguistique (LingwayKM, Kaliwatch, ...)
- la catégorisation des documents collectés : les documents collectés sont affectés à des catégories prédéfinies (Exalead, Kaliwatch, ...)
- l'analyse bibliométrique (Intellixir, Tétralogie, ...)
- les similarités et la visualisation sous forme de réseaux (Lexiquest, Wordmapper, Maspstan, ...)
- la classification automatique et de cartographies (Wordmapper, Tétralogie, ...)
- l'analyse temporelle de l'information (Intellixir, Lexiquest Mine, ...)

Nous nous sommes intéressés plus particulièrement aux outils qui proposent une similarité entre textes. La plupart des outils aborde la similarité à partir d'une indexation (manuelle ou automatique). Des logiciels comme Lexiquest Mine ou WordMapper proposent une similarité entre mots : une classification automatique des termes utilisés dans les textes va être cartographiée et en cliquant sur les mots ou groupes de mots, les textes liés à ces termes sont affichés. D'après Bondu [Bon10], il n'existe pas à l'heure actuelle de logiciel permettant de répondre à notre problématique.

D'après Ihadjadene et al.[IFC03], dans les entreprises deux grands types de solutions sont utilisés pour faire de la veille :

- les solutions intégrées : outils de veille complets censés traiter le processus du début à la fin (collecte, traitement et diffusion de l'information[IFC03],
- les solutions « bricolées » : plusieurs outils traitant l'information à différentes étapes sont combinés pour répondre au besoin des décideurs.

« *Les outils logiciels n'assistent pas la totalité du processus [de veille] mais le complexifient :*

- *parce qu'ils génèrent plutôt une activité nouvelle qui redouble le besoin d'exper-*

tise intellectuelle. Plus l'"intelligence"(au sens anglo-saxon) est technologique, plus elle nécessite de l'assistance humaine, ce qui induit des coûts supplémentaires [...] ;
- *parce que les logiciels "intégrés" nécessitent des compétences très élaborées et variées pour être opérationnels et leur gestion est lourde,*
- *parce que ceux qui ne sont pas intégrés fractionnent l'activité en s'appuyant sur une diversité d'outils qui ne sont interfaçables.*

»[FI04].

4.4 Informations voisines

Nous introduisons le concept d'« information voisine ». « *"Voisine" signifie qu'une information, déjà présente dans la base de données, a une signification proche [d'un] signal faible. Mais il est possible que l'information voisine ait très peu de mots en commun avec le signal faible.* »[LL11]. Dans le cadre de cette thèse, nous précisons que deux informations, au sein d'une base de données se rapportant au même sujet donné (ou ordre du jour), sont dites voisines si on peut les rapprocher selon les trois dimensions suivantes :

- les mots en commun et/ou
- les synonymes (par rapport à un certain dictionnaire de synonymes) et/ou
- les mots cooccurrents (deux mots sont dit cooccurrents lorsqu'ils apparaissent souvent ensemble dans les textes d'un corpus).

Le concept « Informations voisines » est utile lorsqu'il s'agit : de rapprocher deux informations (ou plus), pas nécessairement écrites avec les mêmes mots, dans le but de :

- fiabiliser l'une d'elles (ou plusieurs),
- compléter l'une d'elles (ou plusieurs),
- mettre en évidence une incohérence ou une contradiction entre deux informations,
- faciliter l'interprétation de l'ensemble de ces informations,
- mettre en évidence un prolongement de la problématique que l'on ignorait (peripheral vision),

– d'envisager des liens entre des informations (complémentarité, incohérence, etc.).

Qui est concerné par la recherche d'informations voisines ? Est concerné principalement l'animateur qui assure la gestion de la base de données où sont stockés les FULL texts résultant de l'interrogation des diverses sources surveillées. La hiérarchie aura indiqué à l'animateur la question (ordre du jour) abordée lors de la prochaine séance de travail collectif en vue d'éclairer une éventuelle prise de décision stratégique. La base de données est supposée contenir de possibles signaux faibles. Sachant que les bases de données peuvent être énormes et que la moindre recherche sur l'Internet peut produire plusieurs centaines de FULL texts, on comprend le stress que fait peser, sur l'animateur, la pression de temps exercée par la hiérarchie ainsi que l'anxiété qui en résulte [BR09].

Dans cette thèse, nous souhaitons répondre à une demande de l'équipe veille stratégique ; cette demande concerne l'apport d'une aide à l'animateur pour sélectionner pour la séance de création collective de sens des informations voisines à partir d'une base de textes remplie par les traqueurs. Cette aide doit permettre un gain de temps pour l'animateur : en regroupant les informations voisines et en étiquetant ces regroupements, l'animateur ne devrait plus être obligé de lire tous les FULL texts. Dans le cadre de la recherche d'informations voisines, notre objectif est de proposer un outil modulaire qui apporte à l'animateur, dans sa préparation de séance de création collective de sens, une aide à la lecture des FULL texts. Pour cela, il pourra naviguer dans les textes de sa base. L'outil proposera à l'animateur un classement des textes et un étiquetage. Nous souhaitons développer un outil qui nécessite une assistance humaine minimale, un paramétrage minimal et dont la gestion et l'utilisation soient simples. Les outils de veille existants sont basés sur des techniques de fouille de textes et plus précisément de recherche d'information. Le chapitre suivant présente les techniques existantes pouvant être mises en place pour rechercher des informations voisines.

Références bibliographiques

[AC94] E. Auster et C.W. Choo : How senior managers acquire and use information in environmental scanning. *Information Processing & Management*, 30(5):607–618, 1994. 18

[AFN98] AFNOR : Prestations de veille et prestations de mise en place d'un système de veille. Rapport technique, AFNOR, 1998. 8, 19, 21

[Agu67] F.J. Aguilar : *Scanning the business environment.* Macmillan New York, 1967. 17, 18

[Ans75] H. I. Ansoff : Managing strategic surprise by response to weak signals. *California Management Review*, 18(2):21–33, 1975. 11

[Arg96] C. Argyris : Actionable knowledge : Design causality in the service of consequential theory. *The Journal of Applied Behavioral Science*, 32(4): 390, 1996. 8

[Bon10] J. Bondu : Benchmarking des plateformes de veille choisir son outil. Rapport technique, Inter-Ligere et SerdaLab, 2010. 39

[BR09] D. Bawden et L. Robinson : The dark side of information : overload, anxiety and other paradoxes and pathologies. *Journal of Information Science*, 35(2):180–191, avril 2009. 41

[CA93] C.W. Choo et E. Auster : Environmental scanning : acquisition and use of information by managers. *Annual Review of Information Science and Technology*, 28:279–314, 1993. 18

[CFF03] M.-L. Caron-Fasan et A. Farastier : *Veille stratégique et gestion des connaissances*, pages 237–266. 2003. 29, 30, 31, 32, 33

[Cul83] M.J. Culnan : Environmental scanning : The effects of task complexity and source accessibility on information gathering behavior. *Decision Sciences*, 14(2):194–206, 1983. 18

[Dav00] A. David : La recherche intervention, un cadre général pour les sciences de gestion. *In IX conférence internationale de management stratégique, Montpellier*, pages 24–26, 2000. 8

[Dou05] H. Dou : Veille technologique en formulation. *Techniques de l'ingénieur. Génie des procédés*, (J2260), 2005. 16

[DSP88] R.L. Daft, J. Sormunen et D. Parks : Chief executive scanning, environmental characteristics, and company performance : An empirical study. *Strategic Management Journal*, 9(2):123–139, 1988. 18

[DW84] R.L. Daft et K.E. Weick : Toward a model of organizations as interpretation systems. *Academy of management review*, pages 284–295, 1984. 18, 28

[Ele97] D.S. Elenkov : Strategic uncertainty and environmental scanning : the case for institutional influences on scanning behavior. *Strategic Management Journal*, 18(4):287–302, 1997. 18

[EM04] M.J. Eppler et J. Mengis : The concept of information overload : A review of literature from organization science, accounting, marketing, mis, and related disciplines. *The information society*, 20(5):325–344, 2004. 9, 23, 36

[FB03] A. Farastier et B. Ballaz : Le management des connaissances et la fonction achats : la dimension interorganisationnelle et le rôle clé du système d'information. *Cahier de recherche du CERAG*, 2003. 30

[FHH84] J.L. Farh, R.C. Hoffman et W.H. Hegarty : Assessing environmental scanning at the subunit level : A multitrait-multimethod analysis. *Decision Sciences*, 15(2):197–220, 1984. 18

[FI04] L. Favier et M. Ihadjadene : *Les outils de veille et d'intelligence économique*, chapitre 10. Hermes Science Publications, mar 2004. 19, 40

[Fra05] C. François : L'analyse de l'information proposée par les outils de veille. *Regards sur l'IE : le magazine de l'intelligence économique*, 11:60–63, Oct 2005. 37, 39

[Gho88] S. Ghoshal : Environmental scanning in korean firms : organizational isomorphism in action. *Journal of International Business Studies*, pages 69–86, 1988. 18

[Gue09] M. Guechtouli : Comment organiser son système de veille stratégique ? *In Symposium ATELIS*, 2009. 19, 20

[Ham81] D.C. Hambrick : Specialization of environmental scanning activities among upper level executives. *Journal of Management Studies*, 18(3): 299–320, 1981. 18

[Ham82] D.C. Hambrick : Environmental scanning and organizational strategy. *Strategic Management Journal*, 3(2):159–174, 1982. 18

[Hem09] P. Hemp : Death by information overload. *Harvard Business Review*, 87(9):82–89, 121, septembre 2009. PMID : 19736853. 37

[Huf90] A.S. Huff : *Mapping strategic thought*. John Wiley & Sons, 1990. 29

[IFC03] M. Ihadjadene, L. Favier et S. Chaudiron : L'intelligence économique sur internet : évaluation des pratiques en france. *Présenté à Intelligence Economique : Recherches et Applications*, 2003. 39

[JMFL06] R. Janissek-Muniz, H. Freitas et H. Lesca : Veille anticipative stratégique, intelligence collective (vas-ic) : usage innovant du site web pour la provocation d'informations d'origine terrain. *La Revue des Sciences de Gestion, Direction et Gestion*, (218):19–30, 2006. 17

[KC05] S. Kamoun-Chouk : *Veille Anticipative Stratégique : Processus d'Attention à l'Environnement Application à des PMI tunisiennes*. Thèse de doctorat, Université Pierre Mendès France (Grenoble), 2005. 11, 17, 18, 29

[KEN05] Y. KENGEN : Métiers de veille : survivre à la technologie. la veille, les outils froids et la norme afnor xp x50-053. *In Veille stratégique, scientifique et technologique*, 2005. 19

[LD10] H. Lesca et R. Dourai : Traque et remontée des informations de veille stratégique anticipative : une approche par la notion d'épanouissement de soi. *FACEF Pesquisa*, 7(2), 2010. 30

[LE86] R.T. Lenz et J.L. Engledow : Environmental analysis : The applicability of current theory. *Strategic Management Journal*, 7(4):329–346, 1986. 18

[Les86] H. Lesca : *Système d'information pour le management stratégique de l'entreprise : l'entreprise intelligente*. McGraw-Hill, 1986. 8

[Les92] H. Lesca : Le problème crucial de la veille stratégique : la construction du "puzzle". 1992. 25

[Les03a] H. Lesca : *Veille stratégique : la méthode L.E.SCAnning*. Gestion en liberté, ISSN 1625-3132. EEd. EMS, Colombelles, 2003. 9, 17, 19, 20, 21, 22, 23, 24, 28

[Les03b] N. Lesca : La veille stratégique : vers un système d'information pour le management stratégique des discontinuités. *In Présent et futurs des systèmes d'information*. PUG, aug 2003. 19

[Les08] M.-L. Lesca, N. et Caron-Fasan : Facteurs d'échec et d'abandon d'un projet de veille stratégique : retour d'expériences. *Systèmes d'Information et Management*, 13(3), 2008. 37

[Les10] H. Lesca : *Chimie durable et signaux faibles : le cas du CO2 vu comme une matière à valoriser*, chapitre 110. Traités IC2. Série Technologies et développement durable. Lavoisier, Paris : Hermès science publications, 2010. 34, 35

[LKC09] H. Lesca, S. Kriaa et A. Casagrande : Veille stratégique : Un facteur d'échec paradoxal largement avéré : la surinformation causée par l'internet. cas concrets, retours d'expérience et piste de solutions. *La revue des sciences de gestion*, (245-246):35–42, 2009. 36

[LL10] H. Lesca et E. Lesca : *Gestion de l'Information*. EMS, 2010. 10

[LL11] H. Lesca et N. Lesca : *Les signaux faibles et la veille anticipative pour les décideurs : Méthodes et applications*. Hermes Science Publications, mai 2011. 10, 11, 12, 14, 16, 21, 24, 25, 28, 34, 40

[ML10] S. Medhaffer et H. Lesca : *L'animation de la veille stratégique*. Hermes Science Publications, février 2010. 8, 24, 33

[MSJS00] R.C. May, W.H. Stewart Jr et R. Sweo : Environmental scanning behavior in a transitional economy : evidence from russia. *Academy of Management Journal*, pages 403–427, 2000. 18

[MW04] A.V. Medeiros Wanderley : *Conception et implantation d'un système d'intelligence compétitive dans une entreprise pétrolière dans un environnement de déréglementation*. Thèse de doctorat, Université Paul Cézanne (Aix-Marseille), 2004. 16

[NK98] I. Nonaka et N. Konno : The concept of" ba" : Building a foundation for knowledge creation. *California Management Review*, 40(3):40–54, 1998. 29, 31, 32

[Non91] I. Nonaka : The knowledge-creating company. *Harvard Business Review*, 69(6):96–104, 1991. 30

[NT95] I. Nonaka et H. Takeuchi : *The knowledge-creating company : How Japanese companies create the dynamics of innovation*. Oxford University Press, USA, 1995. 30

[NT03] I. Nonaka et R. Toyama : The knowledge-creating theory revisited : knowledge creation as a synthesizing process. *Knowledge Management Research & Practice*, 1(1):2–10, 2003. 28, 29, 32

[PS78] J. Pfeffer et G.R. Salancik : *The external control of organizations : A resource dependence perspective*. Harper and Row, 1978. 18

[Rab] F. Rabat : Petite contribution à une définition opératoire du concept d'information. http ://docsdocs.free.fr/spip.php ?article374. 9

[SA97] M. Salles et A.-M. Alquier : Réflexions méthodologiques pour la conception de systèmes d'intelligence économique de l'entreprise. *actes du Congrès international " le Génie Industriel dans un monde sans frontières*, pages 3–5, 1997. 19

[SFN88] L.R. Smeltzer, G.L. Fann et V.N. Nikolaisen : Environmental scanning practices in small business. *Journal of Small Business Management*, 26(3):55–62, 1988. 18

[SM99] O.O. Sawyerr et J.E. McGee : The impact of personal network characteristics on perceived environmental uncertainty : an examination of owners/managers of new high technology firms. *Frontiers of Entrepreneurship Research*, 1999. 18

[Tho03] J.D. Thompson : *Organizations in action : Social science bases of administrative theory*. Transaction Pub, 2003. 18

[Wei79] K.E. Weick : *The social psychology of organizing*, volume 2. Addison-Wesley, 1979. 29

[YLBH10] A. Yurchyshyna, M. Léonard et P. Brough-Heinzman : Towards a services-based approach for supporting idea development process. *In Proceedings of the 2010 Fifth International Conference on Internet and Web Applications and Services*, ICIW '10, pages 321–326, Washington, DC, USA, 2010. IEEE Computer Society. 8

Chapitre II

État de l'art en recherche d'information

Pour répondre à la problématique de la veille stratégique, nous nous sommes intéressés aux recherches menées dans le domaine de la recherche d'informations. Nous présentons dans ce chapitre les deux méthodes qui peuvent permettre à l'animateur de trouver de l'information pour la séance de création collective de sens : la recherche et la navigation. Ensuite, nous montrons que les outils de recherche d'information reposent sur 5 fonctionnalités : filtrage (filtre l'information), abstraction (représentation de l'information), lissage (extension de l'abstraction), mesure (comparaison des abstractions) et visualisation des résultats. Nous terminons ce chapitre en présentant la notion de pertinence qui est cruciale en recherche d'information et les méthodes d'évaluation de cette pertinence pour les outils de recherche d'information : la précision et le rappel.

Sommaire

La problématique rencontrée en veille stratégique s'inscrit pleinement dans le domaine de la fouille de textes (voir figure II.1) qui fait partie du domaine plus large de la découverte des connaissances à partir des textes (DCT)[IS07]. Kodratteff cité par Ibekwe [IS07] « *définit la DCT comme "la science qui découvre les connaissances dans les textes". [...] "les connaissances découvertes doivent être ancrées dans le monde réel et doivent modifier le comportement d'un agent humain ou mécanique"* ». Plus précisément, nous nous sommes intéressés à la recherche d'information. Le terme « recherche d'information »(RI) ou « information retrieval » est employé pour la première fois par Moers [Moo48] pour désigner le processus d'indexation automatique et de recherche d'information. Les premiers projets de RI portaient sur l'indexation de documents (projet Cranfield, projet MEDLARS, SMART, ...). Depuis l'avènement d'Internet et par conséquent l'explosion de l'information disponible, la RI s'est vue confrontée à de nouveaux problèmes comme par exemple la surabondance d'information, la redondance, le problème de la qualification de l'information... De plus, la recherche d'information ne concerne plus seulement la documentation ; des techniques de recherche d'information apparaissent dans de nombreux domaines tels que l'analyse de données, la bio-informatique, la linguistique, les statistiques, l'optimisation de grandes bases de données, l'intelligence artificiel... La grande variété des méthodes souligne la diversité des communautés qui travaillent sur le domaine de la recherche d'information.
Pour Baeza-Yates et Ribeiro-Neto [BYRN99], la RI traite de la représentation, du stockage, de l'organisation et de l'accès à l'information. L'utilisateur doit pouvoir accéder facilement à l'information qui l'intéresse. En psychologie, la recherche d'information est considérée « *comme une alternative à la résolution de problème, autre activité mise en œuvre par l'être humain quand il manque de connaissance pour réaliser une tâche* »[TR04]. En effet, le besoin d'information résulte d'un manque de connaissances que l'être humain va chercher à combler. Le dispositif de veille stratégique a pour but d'aider les dirigeants à prendre leurs décisions. Pour la partie qui nous concerne, l'animateur a besoin d'informations pour préparer la séance de création collective de sens et va utiliser un outil de recherche d'information. D'après Baeza-Yates et Ribeiro-Neto [BYRN99] et Sacco et Tzitzikas [ST09], afin d'atteindre son objectif, l'animateur peut procéder de deux façons en réalisant soit une recherche (« retrieval » [BYRN99] ou « focalized search » [ST09]) soit une explo-

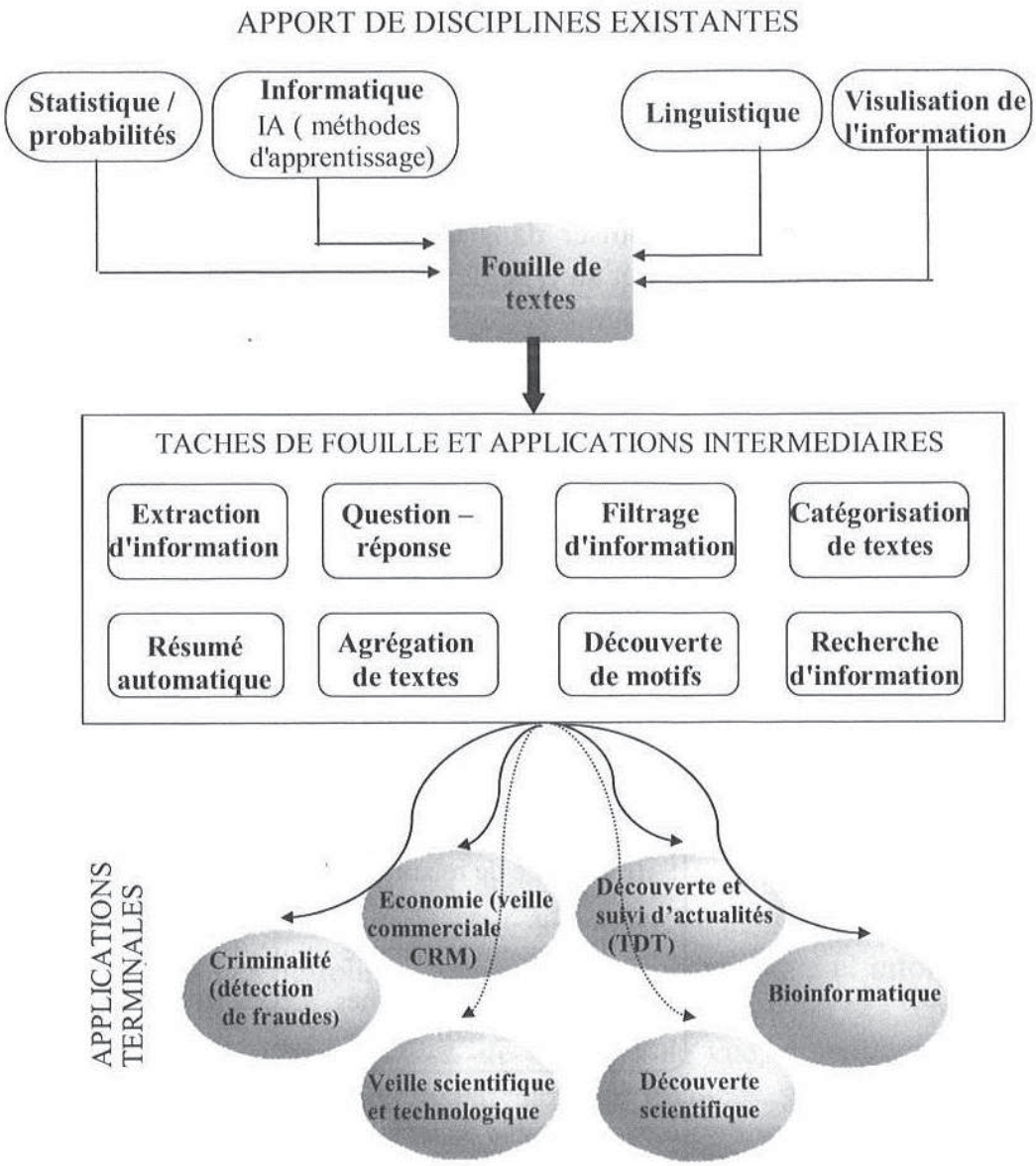


Figure II.1 – Apports disciplinaires et domaines d'application de la fouille de textes [IS07]

ration/navigation (« browsing » [BYRN99] ou « exploratory search » [ST09]). Nous présentons dans une première partie la recherche de documents et la navigation dans des documents. Nous montrons ensuite que les outils de recherche ou de navigation présentent jusqu'à 5 fonctionnalités pour répondre aux besoins de l'utilisateur. Nous terminons ce chapitre par une présentation des méthodes d'évaluation en recherche

d'informations.

1 Recherche et navigation

1.1 Recherche

Dans la recherche, l'utilisateur doit traduire son besoin d'information en mots-clefs ou requêtes nécessaires à l'utilisation d'un système de recherche d'information (SRI).« *Le but des SRI est de représenter et stocker une collection de documents textuels de différents types et tailles, tels que livres, articles de journaux ou rapports techniques, de manière à ce qu'ils soient facilement accessibles par un utilisateur* »[Loi04]. Le SRI va comparer la requête de l'utilisateur aux documents et les documents jugés pertinents seront fournis à l'utilisateur. Clinchant et Gaussier[GY11] indiquent qu'un système de recherche d'information est constitué de trois modules :

- un module d'indexation des requêtes,
- un module d'indexation des documents,
- un module d'appariement entre documents et requêtes.

Pour Boughanem dans [BS08], « *l'indexation a pour rôle d'extraire à partir d'un document ou d'une requête, une représentation paramétrée qui couvre au mieux son contenu sémantique. Le résultat de l'indexation constitue le descripteur du document ou de la requête, qui est une liste de termes significatifs* ». Des analyses sont effectuées pour choisir l'ensemble des termes pertinents pour décrire les documents. Nous reviendrons plus précisément sur l'indexation dans la partie 2.1.
Nous présentons ici les trois principaux modèles de systèmes de recherche d'informations.

Le modèle booléen Dans ce modèle, chaque document D_i est représenté par un ensemble de descripteurs $\{d_1, .., d_j, ..., d_n\}$. Tous les descripteurs des documents sont rangés dans un fichier appelé dictionnaire. Une requête est composée d'un ensemble de descripteurs et un ou des opérateurs logiques comme « ET », « OU » ou « NON ». Par exemple, je recherche un document sur le modèle booléen en recherche d'information, ma requête pourra s'écrire : « modèle ET booléen ET recherche ET

information ». Le système évalue chaque document en fonction de la requête : ainsi tous les documents dont la liste des descripteurs correspond à la requête seront fournis à l'utilisateur. Sur notre exemple, tous les documents ayant exactement dans leur liste « modèle », « booléen », « recherche » et « information » nous seront présentés. Les documents auxquels il manquerait un descripteur ne seront pas fournis.
Les principaux avantages de ce modèle sont :

- sa transparence : l'outil restitue les documents qui répondent exactement à la requête de l'utilisateur,
- sa facilité de mise en œuvre.

Il présente néanmoins des limites :

- la nécessité d'une bonne maîtrise des opérateurs pour obtenir exactement ce que l'on cherche,
- les documents ne sont pas classés et leur nombre pas maîtrisé,
- un document qui ne correspond pas à la requête sur un seul point sera rejeté.

Pour remédier à ces inconvénients, des extensions ont été proposées. Elles ajoutent un poids aux descripteurs et cette pondération sera prise en compte pour calculer un degré de pertinence vis-à-vis de la requête. On peut citer par exemple le modèle booléen étendu[SFW83] ou encore le modèle basé sur la logique floue [Loi04][Iha04].

Le modèle vectoriel Contrairement au modèle booléen, l'utilisateur n'a pas besoin d'exprimer sa requête à l'aide d'opérateurs. Les documents et les requêtes sont représentés par des vecteurs : à chaque composante du vecteur est associé un descripteur issu de l'indexation. La valeur de la composante est le poids attribué au descripteur par rapport au document. Le modèle le plus simple est :

- on met la composante à 1 si le descripteur est attribué au document,
- 0 sinon.

« *Un document est d'autant plus pertinent à une requête que le vecteur associé est similaire à celui de la requête* »[BS08]. Ainsi, on va utiliser un calcul de similarité pour obtenir une liste ordonnée de documents pertinents. Dans la partie 2, nous présentons une autre attribution de poids pour les composantes ainsi que des mesures de similarité.

Les modèles probabilistes Le principe des modèles probabilistes est de présenter les résultats d'une recherche en fonction de la probabilité de pertinence d'un document par rapport à une requête. Selon Clinchant et Gaussier [CG11], il existe trois classes de modèles probabilistes :

- Principe d'ordonnancement probabiliste : « *ces modèles supposent que pour une requête il existe une classe de documents pertinents et une classe de documents non pertinents. Cette idée conduit à ordonner les documents selon la probabilité de pertinence du document* »[CG11],
- Modèles de langue : la pertinence d'un document par rapport à une requête est donnée par la probabilité que le document génère la requête (modèle le plus connu et le plus utilisé actuellement),
- Approches informationnelles : « *ces approches visent à quantifier l'importance d'un terme dans un document par rapport à son comportement dans la collection* »[CG11]. Par exemple, un mot fréquent dans un document mais qui apparaît dans peu de documents sera très important pour caractériser ce document.

Actuellement, l'animateur utilise les SRI et doit formuler son besoin sous forme de requêtes. Comme on l'a vu au chapitre I, ce travail est long et fastidieux. De plus, la transformation en mots-clefs de son besoin d'information n'est pas évidente et le choix de ces mots a une influence sur les résultats. L'animateur a ainsi de plus grands risques de passer à côté d'informations intéressantes. Nous pensons donc que la navigation est plus adaptée au besoin de l'animateur de veille stratégique.

1.2 Navigation

Pour Toms [Tom00], « *the definition of browsing combines both concepts to define browsing as an activity in which one gathers information while scanning an information space without an explicit objective* ». L'utilisateur n'a pas besoin pour la navigation de traduire son besoin d'information en requête. Il « navigue » dans les documents pour répondre à ses besoins qui peuvent évoluer au fur et à mesure de la consultation des documents.
Par rapport aux problèmatiques de la veille stratégique, nous pensons que la navi-

gation présente de meilleurs atouts que la recherche. « *In the case of textual information, not all information is equally useful or pertinent to each person ; the set of experiences and knowledge of the user at the time of the interaction in combination with the conceptual properties of the text influences which textual artifact the user will activate and, thus, the information that will be gathered. Accordingly, browsing relies on the experiences and personal attributes of the browser (i.e. the "prepared mind") in combination with the topography of the text as presented by the system (which together form those textual affordances)* »[Tom00]. Le terme « browsing » est traduit en français soit par navigation soit par butinage. Le Crosnier[LC91] distingue la navigation du butinage : « *le terme de navigation à la circulation utilisant une carte de navigation. Cette activité représente une action réfléchie et contrôlée à partir d'un projet général, ayant une destination particulière.[...] [Le butinage] s'apparente plus à la flânerie au sein de l'univers informationnel, et constitue une activité cognitive plus difficile à modéliser* ». Nous nous plaçons dans le cadre d'une navigation. Nous présentons ici les différents types de « browsing »[BYRN99] :

Navigation plate La navigation plate [BYRN99] propose à l'utilisateur de parcourir un document ou une liste de documents sans organisation particulière. L'utilisateur navigue dans le document et s'arrête sur les points qui l'intéressent. Il peut ainsi explorer des documents qui vont l'aider à formaliser sa recherche en lui suggérant des mots clefs. Pour naviguer, l'utilisateur utilise par exemple la barre de navigation.

Navigation structurée Pour faciliter la navigation, les documents sont présentés à l'utilisateur sous une forme structurée. Il peut s'agir par exemple des documents rangés dans des dossiers et sous-dossiers [BYRN99]. L'organisation des documents peut être hiérarchique ou sous forme de graphe.

Navigation par lien hypertexte « *Sera désigné comme hyperdocument tout contenu informatif informatisé dont la caractéristique principale est de ne pas être assujetti à une lecture préalablement définie mais de permettre un ensemble plus ou moins complexe, plus ou moins divers, plus ou moins personnalisé de lectures. Parcourant des hyperdocuments, le lecteur peut, dans une certaine mesure, décider de sa lecture et agir sur elle en définissant ses parcours. Un hyperdocument est donc tout*

contenu informatif constitué d'une nébuleuse de fragments dont le sens se construit, au moyen d'outils informatiques, à travers chacun des parcours que la lecture détermine »[Bal90]. Un hyperdocument est composé de noeuds liés entre eux par des hyperliens qui permettent de passer automatiquement d'un nœud à un autre. Un noeud est une unité minimale d'information : par exemple du texte, un élément audio, une vidéo, ... Aujourd'hui le web est le système hypertexte le plus vaste et le plus utilisé au monde. Par simple clic, on navigue d'un document à un autre.
Les avantages de ce type de navigation sont :

- un accès rapide à une grande quantité d'information,
- un accès plus ou moins détaillé selon son besoin : si l'on souhaite plus d'informations, on clique sur l'hyperlien qui nous permet d'accéder à une information plus détaillée.

Ses inconvénients sont :

- les difficultés d'orientation : la multiplicité de liens peut perdre l'utilisateur,
- le maintien de la cohérence des liens.

2 Fonctionnalités des outils de recherche d'information

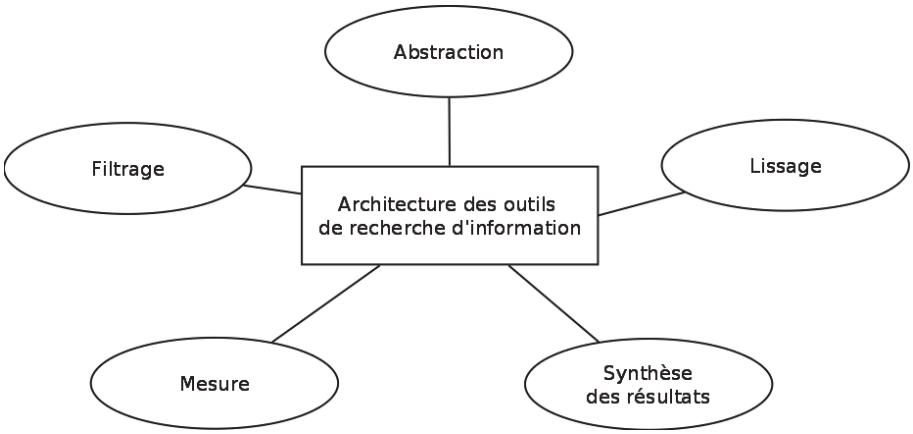


Figure II.2 – Les fonctionnalités des outils de recherche d'information

Après une étude des outils de recherche et de navigation et de la littérature associée, nous proposons une grille de lecture de ces outils en cinq fonctionnalités

(voir figure II.2) :

- abstraction
- filtrage
- lissage
- mesure
- visualisation des résultats.

2.1 Abstraction

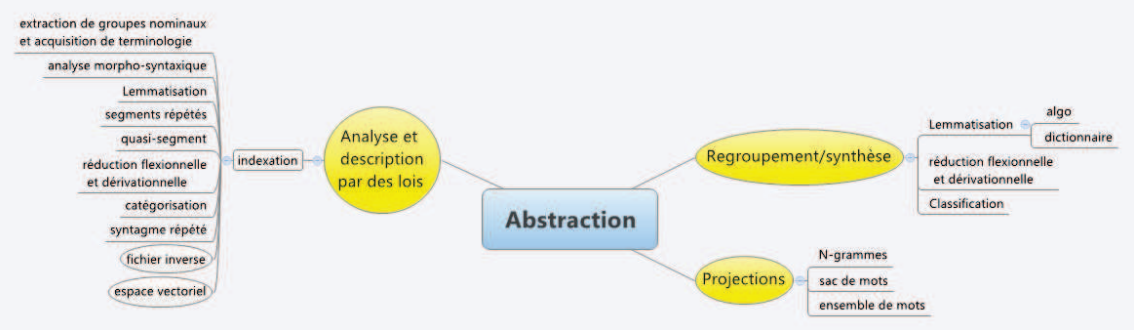


Figure II.3 – Abstraction

Afin de comparer automatiquement des textes, il est nécessaire de les rendre exploitables par l'ordinateur. Les outils de recherche d'information créent une représentation plus abstraite d'un texte : des informations différentes vont être représentées de la même manière. L'avantage de cette fonctionnalité est d'assurer une représentation homogène à partir de données hétérogènes. En revanche, cette abstraction se fait au détriment de la richesse d'un texte : cela se traduit par une perte d'informations. « L'abstraction finale » peut être la composition de plusieurs abstractions.
En s'inspirant de Pincemin[Pin99], les abstractions peuvent être réparties en trois groupes :

- les projections
- les regroupements/synthèses
- Les analyses et descriptions par des lois

Projections « *La projection est la réification de la perception : elle explicite ce que*

capte la machine, ce qu'aperçoit le lecteur »[Pin99].« *Pour un texte : va-t-on considérer la police de caractère utilisée, l'ordre des mots, le changement de page... Le choix le plus courant, en lexicométrie, est de considérer le texte comme une suite de caractères, en distinguant des caractères constitutifs (lettres), des caractères séparateurs de mots (espace, apostrophe), et éventuellement des caractères particuliers (ponctuation)* »[Pin99]. En recherche d'information, l'approche classique repose sur la notion de « sac de mots » ou « ensemble de mots » [BYRN99][Tri07]. Les aspects concernant l'ordre des mots, la structure des phrases sont supprimés pour ne conserver que les mots. Ainsi deux textes ayant exactement les mêmes mots mais dont l'ordre est différent seront considérés comme identiques.
Mellet et Barthélemy[MB09] estiment qu'un texte ne doit pas être considéré comme sac de mots mais « *qu'un texte est une structure linéaire constituée d'un ensemble d'événements linguistiques (occurrences d'un mot, d'un syntagme, d'une catégorie grammaticale, etc.* ». Ces auteurs parlent de « topologie textuelle ». Cette abstraction permet la recherche de motif dans les textes.

Une autre projection possible est de découper les textes en n-grammes[Pin99] : un texte est représenté par un ensemble de suites de n caractères. Par exemple, les 3-grammes de « La représentation » sont _La, La_, _re, rep, epr, ... Les avantages de cette méthode sont :

- qu'un petit nombre de lettres peut représenter tous les mots d'une langue
- qu'elle peut être utilisée dans toutes les langues car elle ne nécessite pas de lexique, de dictionnaire ou de grammaire.

A noter que pour les langues ayant des origines apparentées, certaines racines peuvent se retrouver d'une langue à l'autre. Cependant, la définition purement formelle des n-grammes (suites de n lettres) rend obscure l'interprétation de ces n-grammes. Il est difficile de lier un n-gramme à un ensemble sémantiquement cohérent de mots. De plus, cette abstraction ne se combine avec aucun traitement linguistique.

Regroupements/synthèses« *Le regroupement procède par fusion : ce qui était distingué, et constituait plusieurs unités, est finalement saisi en seule unité. Une information sur la nature de la fusion peut enrichir la nouvelle unité. L'unité joue*

alors le rôle d'une formulation synthétique de ses composantes »[Pin99].
Un regroupement courant utilisé est le regroupement de mots par lemmatisation. La lemmatisation est une opération linguistique qui ramène les formes fléchies (conjuguées, plurielles, féminines) à une forme standard (infinitif, singulier, masculin)[Mel01]. On ne conserve pour les calculs que les formes standards. Les temps de calculs et la mémoire sont ainsi optimisés. Il existe deux types de lemmatisation :

- lemmatisation par algorithme : un algorithme regroupe les mots d'une même famille comme par exemple transformer, transformation, transforment... Cette méthode nécessite une connaissance de la langue et de la manière dont les mots sont formés. Porter [Por97] propose un algorithme qui en 5 étapes élimine les terminaisons des mots en anglais. Il existe d'autres algorithmes : sample text, Lovins stemmer, Paice stemmers, ... [MRS08]
- lemmatisation par dictionnaire [Nie08][Sav93] : pour savoir si une séquence de lettres à la fin correspond à une terminaison, il suffit de faire une élimination ou une transformation tentative, et de voir si la forme obtenue existe dans le dictionnaire. Sinon, ce n'est pas une terminaison correcte, et d'autres possibilités sont ensuite envisagées. Par exemple, on peut accepter la règle qui remplace -ation par -er .

Le principal inconvénient de la lemmatisation est la perte d'information lorsqu'on utilise un lemme à la place du mot.« *Le danger de la lemmatisation est là : avec elle, nous nous imaginons sauvés de la noyade linguistique, alors même que, se croyant à l'abri, l'on reste dangereusement au milieu du guets. "La lemmatisation ne résout rien et empire tout" (Tournier, 1985 : 485) conclut Maurice Tournier non sans pertinence : elle ne résout pas la question du sens - indécidable hors contexte et surtout pas par une machine - et travestit déjà le texte effectivement émis que l'on se proposait d'étudier* »[May05]. Dans le cadre d'un système de questions-réponses, Billoti et al [BKL04] observent de moins bons résultats une fois leur corpus lemmatisé. Lemaire [Lem08] note également une perte de performance lors de l'étude du contexte d'apparition des mots.
La réduction flexionnelle et dérivationelle [Pin99] est une autre possibilité pour abstraire les mots d'un texte. On dissocie pour chaque mot ses parties lexicales(racine, affixes) et ses parties grammaticales (déclinaison, conjugaison). Les mots sont regroupés par famille en choisissant de supprimer les composantes ajoutées à la racine. Par

exemple, « expérience », « expérimental », « expérimenter », « expérimentalement » seront regroupés ainsi que « expert », « expertise ». Cette méthode permet de récupérer la notion ou le concept sous-jacent et de diminuer le volume de données à traiter. Cependant, tous les rapprochements ne sont pas faits et de mauvais rapprochements sont réalisés : par exemple « courir » et « courant » (électrique) peuvent être rapprochés.
Dans ces méthodes de regroupements, le traitement des textes syntaxiquement ou sémantiquement mauvais reste un problème.

Analyses et descriptions par des lois « *Un ensemble d'unités "primitives", muni de lois de composition, permet de représenter en puissance un beaucoup plus grand nombre d'unités "complexes". Les primitives et les lois constituent un résumé d'un ensemble de possibilités virtuelles* »[Pin99].En recherche d'information, l'abstraction « analyse et description par des lois » la plus utilisée est l'indexation. L'indexation consiste à représenter un texte par un ensemble de descripteurs. Les descripteurs peuvent être des mots, des groupes de mots, des concepts d'une ontologie... On peut distinguer 3 types d'indexation[Baz05][BS08] :

- l'indexation manuelle : un groupe d'experts ou documentalistes analyse les documents et affecte des descripteurs à chacun,
- l'indexation automatique : l'affectation des descripteurs est réalisée par un programme informatique. Les textes sont analysés par un ordinateur et des termes en sont extraits ou sont affectés à partir d'une ressource lexicale telle que Wordnet, un thésaurus, une ontologie... [Baz05] [Her06] :
- l'indexation semi-automatique : une première analyse est réalisée automatiquement. En s'appuyant sur cette analyse, le choix définitif de l'indexation est réalisé par un ou plusieurs experts.

L'extraction de termes/formes pertinent(e)s s'appuie sur plusieurs types d'analyses :

- la lemmatisation
- la réduction flexionnelle et dérivationnelle
- l'extraction des groupes nominaux [Pin99] : s'appuyer sur des groupes nominaux plutôt que de termes seuls se traduit par un gain sémantique car on a des entités qui décrivent de manière plus précise un texte. En revanche, cela entraîne une multiplication des unités et il est donc nécessaire de faire appel

à une personne connaissant le domaine pour choisir les unités pertinentes.
- l'analyse morpho-syntaxique [Sid02] : il s'agit ici d'analyser la forme d'une part et la fonction d'autre part. Grâce à cette analyse, on obtient un nombre de descripteurs peu élevé mais on efface des informations liées au contexte et on a un risque d'attribution redondante (par exemple H_2O et eau).
- la catégorisation : cette opération consiste à découper un texte en unités lexicales et à les identifier grammaticalement (nom, verbe, adjectif, ...). Le catégoriseur le plus connu est Brill.
- les segments répétés[SL83] : un algorithme recherche dans les textes les séquences de 2, 3, ..., n mots qui sont répétés à plusieurs reprises.
- les quasi-segments[BP93] : le quasi-segment est une séquence de formes d'une même phrase, non nécessairement contiguës, répétée à plusieurs reprises dans le corpus.
- les syngtames répétés [PPL98] : groupes de mots qui forment une unité ou qui se suivent avec un sens ; contrairement aux segments répétés, on ne s'intéresse qu'au lemme en excluant les mots grammaticaux c'est-à-dire que seuls les adjectifs, noms ou substantifs sont considérés.

L'indexation, pour être utile, amène à la création :

- d'un fichier inverse : « *Après avoir enregistré, pour chaque document, la liste des termes qu'il contient, on crée un fichier inverse qui dresse, pour chaque terme, la liste des documents qui le contiennent* »[Iha04]. Le principal avantage de l'utilisation d'un fichier inverse est sa facilité de mise en œuvre.
- de vecteurs pour chaque document : on crée un espace vectoriel [SM83][Mem00] défini grâce à la sélection des descripteurs. Un document sera représenté par un vecteur où chacune des composantes sera liée à un descripteur. Dans le cas simple, le vecteur est dit booléen car si le descripteur est affecté au document, on mettra un « 1 » à la place de la composante liée à ce descripteur ; dans le cas contraire on mettra un « 0 ». Par exemple, si nos descripteurs sont « espace vectoriel », « descripteur » et « composante », pour la phrase « on crée un espace vectoriel défini grâce à la sélection des descripteurs » notée *Phrase*, on peut créer le vecteur simple :

	espace vectoriel	descripteur	composante
Phrase	(1	1	0)

Afin de tenir compte de propriétés des documents comme la fréquence d'apparition du descripteur dans le document ou encore la taille du document, des méthodes plus élaborées ont été proposées. La méthode classique en recherche d'information est le calcul du Term Frequency-Inverse Document Frequency(TF-IDF [SM83][BS08][MRS08]). Le TF-IDF est le résultat de la multiplication de :

- TF term frequency : $\mathrm{tf_{i,j}} = \frac{n_{i,j}}{\sum_k n_{k,j}}$ où $n_{i,j}$ représente le nombre de fois où le terme t_i apparaît dans le document d_j et $\sum_k n_{k,j}$ correspond au nombre de mots dans le document d_j.
- IDF inverse document frequency : $\mathrm{idf_i} = \log \frac{|D|}{|\{d_j : t_i \in d_j\}|}$ où $|D|$ est le nombre de documents dans le corpus D et $|\{d_j : t_i \in d_j\}|$ le nombre de documents où apparaît t_i.

L'avantage d'abstraire les textes sous forme de vecteurs est la possibilité d'utiliser par la suite toutes les techniques liées aux vecteurs : normes, calcul d'angle, distance euclidienne, ... En revanche, les représentations ne constituent pas « rigoureusement » des espaces vectoriels. Dans la plupart des modèles vectoriels, l'hypothèse d'indépendance des descripteurs est admise. Mais on se rend bien compte qu'en langue naturelle les mots ne sont pas toujours indépendants. Un autre hypothèse de la représentation vectorielle dit qu'un descripteur est une unité de sens pour les documents. Or, on sait que les termes d'une langue sont ambigus et polysémiques.

2.2 Filtrage

Le *filtrage* est une méthode de suppression des informations jugées inutiles et pouvant parasiter les résultats. Il s'agit à cette étape de procéder comme un chercheur d'or qui utilise un tamis pour séparer le sable des pépites. Seuls les mots ou groupes de mots porteurs de sens sont conservés. Le principal avantage de cette fonctionnalité de filtrage est de limiter le bruit, i.e. le nombre de documents non pertinents fournis. En revanche, filtrer l'abstraction augmente le silence, i.e le nombre de documents pertinents non fournis. Cette étape implique une perte d'information.

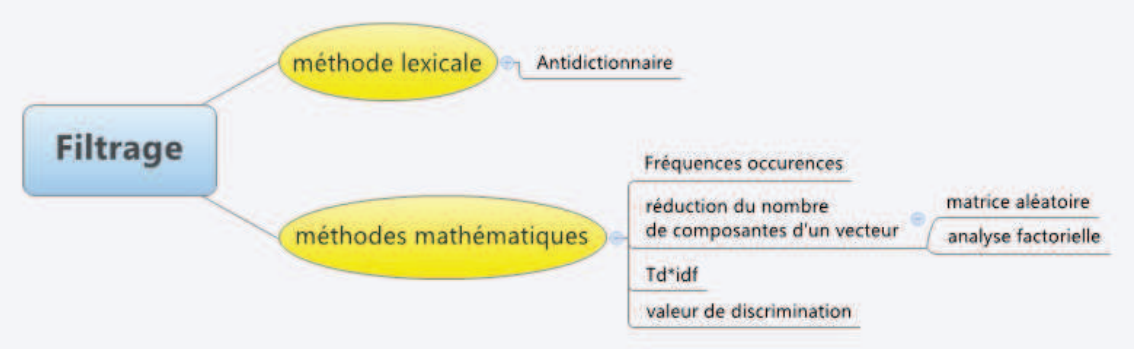


Figure II.4 – Filtrage

Les méthodes de filtrage peuvent être regroupées en deux groupes :

- les méthodes lexicales
- les méthodes mathématiques

Méthodes lexicales Pour filtrer l'abstraction, l'antidictionnaire est très souvent utilisé. Il s'agit de supprimer du texte tous les mots non pertinents tels que les articles (le, la, les, ...), certains verbes (être, avoir, ...) etc. L'antidictionnaire liste l'ensemble des mots que l'on considère comme peu porteurs de sens ou peu discriminants. Ainsi on diminue le nombre d'entités à traiter et on ne conserve que celle jugée pertinente. En revanche « *faire une élimination à partir d'une liste de mots-outils, c'est savoir ce que l'on veut laisser et garder ce dont la preuve de l'inutilité n'est pas (encore) faite* »[Pin99].

Méthodes mathématiques Ces méthodes mathématiques tirent leur origine de la loi de Zipf et la conjecture de Luhn [BS08]. La loi de Zipf prévoit que dans un texte donné, la fréquence d'occurrence $f(n)$ d'un mot est liée à son rang n dans l'ordre des fréquences par une loi de la forme $f(n) = \frac{K}{n}$ où K est une constante. George Kingsley Zipf aurait découvert cette loi en analysant *Ulysses* [Joy09] de James Joyce. En comptant les mots distincts, il aurait remarqué que :

- le mot le plus courant revenait 8 000 fois ;
- le dixième mot 800 fois ;
- le centième, 80 fois ;

– et le millième, 8 fois.

La conjecture de Luhn s'appuie sur la loi de Zipf et émet une hypothèse sur l'information contenue dans les mots (informativité) d'un document schématisée par les deux courbes de fréquence et d'informativité dans la figure[1] II.5. Les mots de rangs

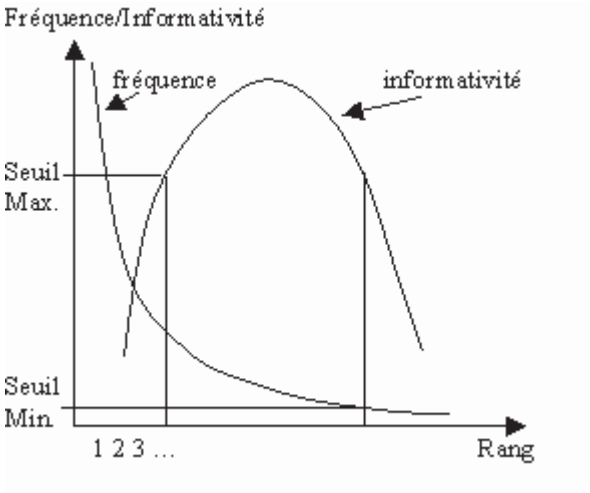


Figure II.5 – Conjecture de Luhn

faibles (i.e. des mots qui reviennent souvent) et les mots de rangs élevés (i.e. des mots rares) sont considérés comme non discriminants. On fixe un seuil minimum et un seuil maximum pour filtrer. Cette approche basée sur la fréquence d'apparition dégage des concepts pour résumer la sémantique d'un texte. Le risque de cette méthode est de faire apparaître comme concepts des mots fonctionnels ou polysémiques.

Des poids peuvent être appliqués à chacun des mots pour rendre compte de leur importance. Le tf-idf est à nouveau très utilisé : on ne conserve que les termes ayant un tf-idf au-dessus d'un seuil fixé.

Dans l'abstraction vectorielle, le nombre de composantes de vecteurs peut être trop important pour les traitements. Les méthodes de réduction des composantes permettent d'obtenir des vecteurs de plus petites tailles avec moins de composantes nulles [Mem00]. On optimise la mémoire et on économise du temps pour les calculs. Néanmoins ces méthodes impliquent une perte d'informations et une complexité des pré-traitements. On peut citer comme exemple :

1. figure tirée du cours de Jian-Yun Nie, http ://www.iro.umontreal.ca/ nie/IFT6255/Indexation.html, consulté le 9 février 2012

– Matrice aléatoire : on multiplie les vecteurs initiaux par une matrice aléatoire et on obtient des vecteurs de dimension moindre [RK89].
– Analyse factorielle : on réalise des analyses factorielles pour calculer la pertinence des termes et n'en retenir qu'un nombre réduit [Mem00].

2.3 Lissage

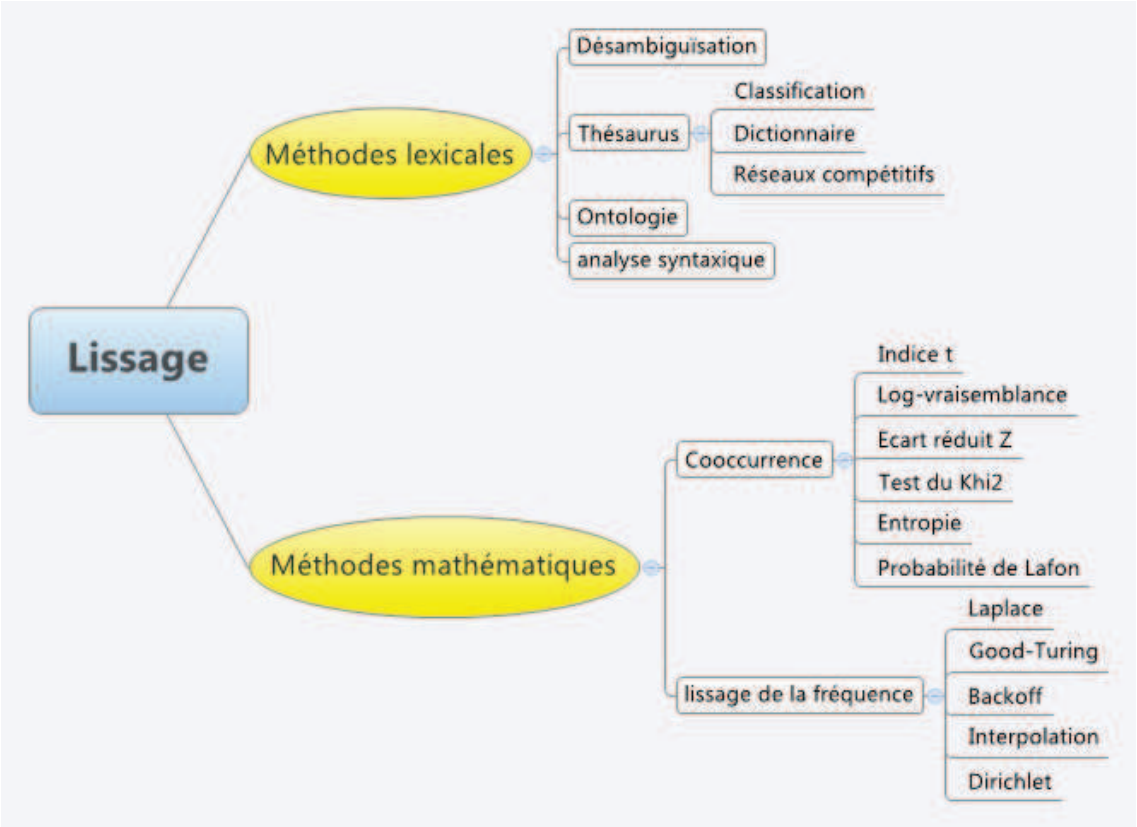


Figure II.6 – Lissage

Le lissage peut être défini comme une méthode d'extension d'un modèle abstrait afin que celui-ci « colle » à d'avantage de résultats dans le contexte d'une recherche. Il s'agit d'une abstraction du modèle initial mais non basée sur des éléments issus du texte. Ainsi, on a un modèle plus large, moins attaché au contexte qui permet de limiter le silence. Néanmoins, cela engendre une perte de finesse du modèle et donc davantage de bruit. Il existe des méthodes lexicales et des méthodes mathématiques :

Méthodes lexicales L'utilisation de thésaurus, de bases de données lexicales telles que Wordnet [MM00] [Voo94] ou d'ontologie [Baz05] permet d'enrichir l'abstraction avec des concepts supplémentaires. Des classes sémantiques de mots sont par exemple créées pour constituer un dictionnaire ou un thésaurus de mots apparentés. Cela permet d'élargir la recherche de documents à des termes auxquels l'utilisateur n'aurait pas pensé[Mem00]. Par exemple, si un texte parle de « vilebrequin », grâce à une ontologie on pourra ajouter le terme « moteur » (voir figure II.7 [2]).

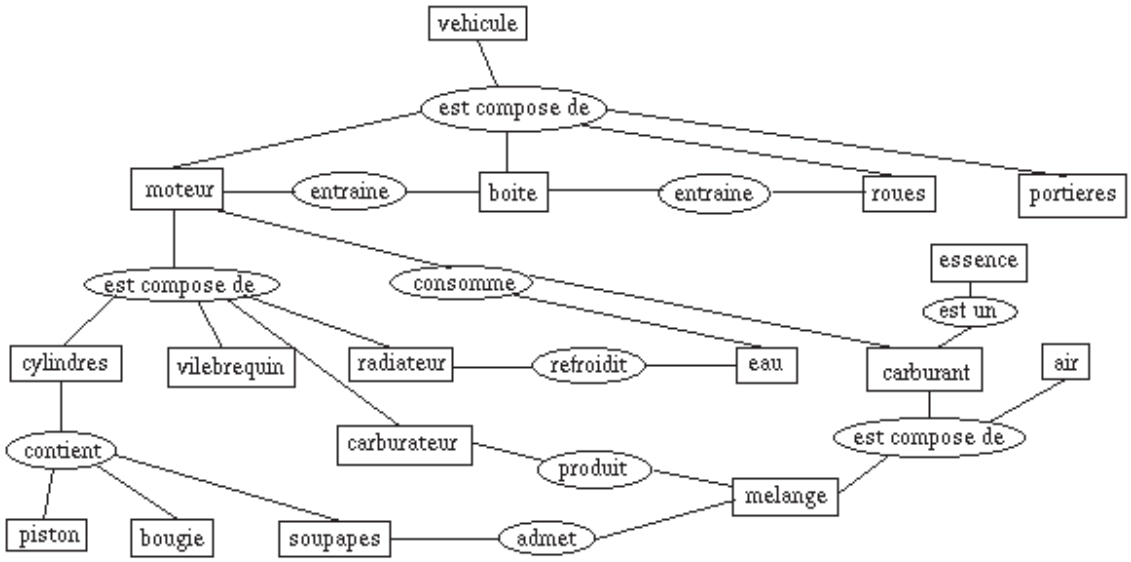


Figure II.7 – Exemple d'ontologie

Une méthode lexicale pour enrichir l'abstraction est la désambiguïsation des termes [IV98]. Les mots en langue naturelle, par exemple le français, sont souvent polysémiques et ambigus. Les algorithmes de désambiguïsation ont pour objectif de préciser le sens d'un mot : par exemple « avocat » peut faire référence au fruit comme à la personne inscrite au barreau. Voorhees [Voo94], Mihalcea [MM00] ont utilisé des données lexicales pour enrichir leur abstraction mais ces auteurs soulignent que cela n'est réalisable que si les termes ont été correctement désambiguïsés.

Des informations syntaxiques et sémantiques peuvent être intégrées à l'abstraction.

2. tirée de http ://liris.cnrs.fr/alain.mille/enseignements/IGC_M2_2008/session8/rapc2/Rapc_Session2_Cas_base_

L'analyse syntaxique apporte une information sur les regroupements et les dépendances entre les unités lexicales et sur les fonctions recensées par la grammaire. Besançon [Bes01] intègre différentes connaissances :

- des connaissances syntaxico-sémantiques : les multi-termes (mots composés ou expressions complexes) sont identifiés à l'aide de patron morpho-syntaxiques tels que Nom+Adjectif (exemple : flux migratoire) ou Adjectif+Nom ou encore Nom + Préposition + Article + Nom + Adjectif (exemple : crainte de la communauté internationale).
- des connaissances sémantiques telles que la synonymie.

Méthodes mathématiques Pour lisser l'abstraction, certains chercheurs utilisent la cooccurrence. Pour Thoiron et Béjoint [TB89], la notion générale de cooccurrence repose sur le fait que, d'une manière générale, les mots ont des affinités particulières. Ces affinités sont essentiellement de deux types :

- affinités seulement sémantiques
- affinités syntactico-sémantiques : les mots entretiennent des relations sémantiques étroites à l'intérieur d'une structure syntaxique.

Il existe de nombreux modèles de cooccurrence [Hei04] ; on peut citer par exemple l'indice t, log vraisemblance, l'écart réduit Z, le test du khi2, l'entropie, le modèle de cooccurrence de Lafon, la distance de Cilibrasi et Vitanyi [CV07] ...
Dans les représentations de textes qui tiennent compte de la fréquence d'apparition des termes dans les textes, il existe plusieurs méthodes pour lisser la fréquence. Boughanem et al. [BKN04a] présentent plusieurs techniques telles que le lissage de Laplace, celui de Good-Turing, lissage Backoff, le lissage par interpolation et le lissage de Dirichlet.

2.4 Mesure

La fonctionnalité « Mesure » va évaluer la similarité entre deux objets : deux documents ou un document et une requête. Qu'est-ce qu'une similarité ? Bisson [Bis00] propose une définition générale des fonctions de similarité :
une fonction de similarité est définie dans un univers $\mathcal{U}$ qui peut être modélisé à

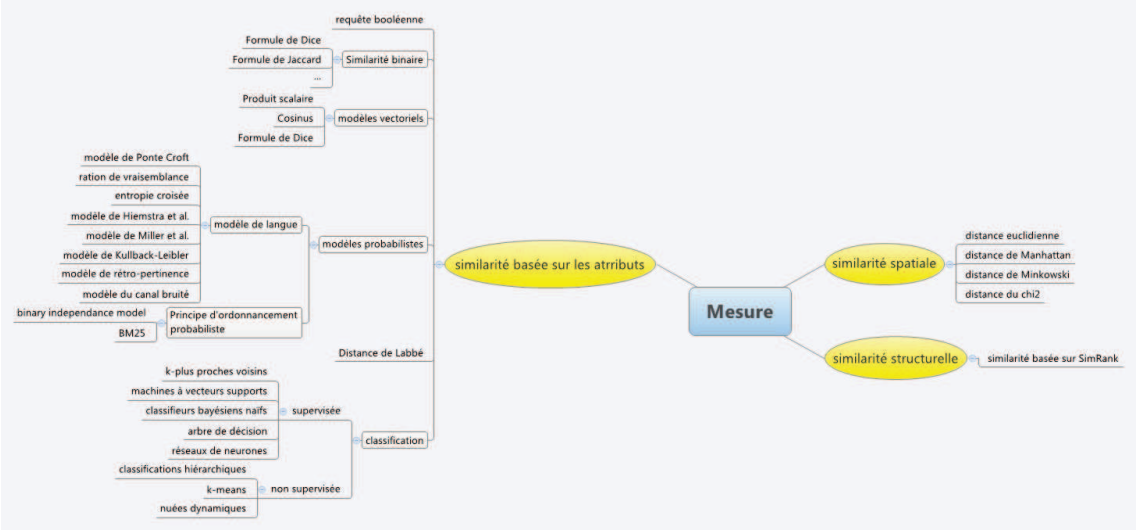


Figure II.8 – mesure

l'aide d'un quadruplet $(L_d, L_s, \mathcal{T}, \mathcal{FS})$.

- Soit L_d le langage de représentation utilisé pour décrire les données.
- Soit L_s le langage de représentation des similarités.
- Soit $\mathcal{T}$ un ensemble de connaissances que l'on possède sur l'univers étudié.
- Soit $\mathcal{FS}$ la fonction de similarité, telle que : $\mathcal{FS} : L_d \times L_d \longrightarrow L_s$

Le principal avantage d'utiliser une fonction de similarité est la possibilité de réaliser des calculs mathématiques qui serviront de base pour la comparaison. Ainsi un texte pourra être jugé plus ou moins proche d'un autre texte à l'aide de ce calcul.
On peut néanmoins se poser la question de la signification de réaliser des calculs à partir de textes. Pour Brunet dans son article « Peut-on mesurer la distance entre deux textes ? »[Bru03], « *certes le texte a une réalité matérielle qui se prête à l'analyse. [...] Pour notre part nous craignons non seulement la parcellisation et l'incohérence des résultats mais aussi le danger de se tromper dans l'interprétation, même quand ces résultats semblent clairs et convergents. Même lorsqu'une distance paraît établie solidement entre deux textes, on se sait pas toujours à quoi la rattacher. A l'auteur ? A l'époque ? Au sujet traité ? Au genre littéraire ?* ». Barthélémy et al. [BLM03] estiment qu'une mesure présente une part d'arbitraire et des biais. En effet la mesure utilisée dépend de l'abstraction choisie.

Nous allons maintenant présenter des exemples de mesures utilisées en recherche d'information. Il en existe tant que notre liste ne sera pas exhaustive mais donnera une idée de la variété existante. Ces diverses approches selon Champclaux et al. [CDM10] s'appuient sur 3 philosophies :

- deux entités sont similaires si elles peuvent être comparées en terme de distance dans un espace qui le permet (similarité spatiale),
- deux entités sont similaires si elles partagent plus de ressemblances que de dissemblances (similarité basée sur les attributs)
- deux entités sont similaires si elles partagent des attributs et des relations entre des attributs similaires (similarité structurelle)

Similarités spatiales En recherche d'information, les principales similarités spatiales sont basées sur l'abstraction vectorielle :
dans la suite, on note :
$V_A = (v_{1,A}, ..., v_{i,A}, ..., v_{n,A})$ et $V_B = (v_{1,B}, ..., v_{i,B}, ..., v_{n,B})$ les vecteurs de deux textes A et B.
On peut citer comme similarité spatiale [RJ08] :

Distance euclidienne	$\sqrt{\sum_{i=1}^{n} \lvert v_{i,A} - v_{i,B} \rvert^2}$	(II.1)
Distance de Manhattan	$\sum_{i=1}^{n} \lvert v_{i,A} - v_{i,B} \rvert$	(II.2)
Distance de Minkowski	$\sqrt[p]{\sum_{i=1}^{n} \lvert v_{i,A} - v_{i,B} \rvert^p}$	(II.3)
Distance du χ^2	« *la distance du χ^2 est la distance euclidienne entre deux vecteurs normalisés par leur longueur (nombre de mots), pondérée par la masse de chacun des mots de l'ensemble des textes (nombre total d'un mot dans l'ensemble des textes)* »[RJ08]	

Similarités basées sur les attributs

Requête booléenne	La première similarité utilisée en recherche d'information est la requête booléenne : la requête booléenne étant formulée à l'aide d'opérateur, elle évalue si le document est conforme à la demande de l'utilisateur.
Similarités binaires	Les similarités binaires sont des mesures structurelles. Ces mesures se basent pour deux textes A et B sur : – a = ce qui est commun à A et B – b = ce qui est dans A mais pas dans B – c = ce qui est dans B mais pas dans A – d = ce qui n'est ni dans A ni dans B Choi et al.[CCT10] recense 76 mesures binaires différentes. Nous présentons ici seulement les deux plus utilisées. **Indice de Jaccard** L'indice de Jaccard vient de l'écologie et il est défini par : $$Jaccard = \frac{a}{a+b+c} \quad \text{(II.4)}$$ Dans le cas d'une abstraction "sac de mots", si on note E_A et E_B les ensembles de mots sélectionnés respectivement des textes A et B, l'indice de Jaccard s'écrit : $$Jaccard = \frac{\mid E_A \cap E_B \mid}{\mid E_A \cup E_B \mid} \quad \text{(II.5)}$$ L'indice de Jaccard représente la proportion de mots communs à A et B sur l'ensemble des mots de A et B.

	Mesure de Dice La mesure de Dice est définie par $$Dice = \frac{2a}{2a+b+c} \quad \text{(II.6)}$$ Dans le cas d'une abstraction "sac de mots", si on note E_A et E_B les ensembles de mots sélectionnés respectivement des textes A et B, la mesure de Dice s'écrit : $$Dice = \frac{2 \mid E_A \cap E_B \mid}{\mid E_A \mid + \mid E_B \mid} \quad \text{(II.7)}$$
Similarités liées au modèle vectoriel	**Produit scalaire** La similarité la plus simple est le produit scalaire.Soit $V_A = (v_{1,A}, ..., v_{i,A}, ..., v_{n,A})$ et $V_B = (v_{1,B}, ..., v_{i,B}, ..., v_{n,B})$ les vecteurs de deux textes A et B, le produit scalaire est : $$V_A \dot{V}_B = \sum_{i=1}^{n} v_{i,A} \cdot v_{i,B} \quad \text{(II.8)}$$ Le produit scalaire mesure l'intersection entre les deux documents c'est-à-dire ce qui est commun **Cosinus** La similarité la plus utilisée est le cosinus : elle mesure l'angle entre les deux vecteurs représentatifs des textes Soit V_A et V_B les vecteurs de deux textes A et B, le cosinus est : $$\cos(V_A, V_B) = \frac{V_A \cdot V_B}{\|V_A\| \|V_B\|} \quad \text{(II.9)}$$ **Mesure de Dice** Dans le cas d'une abstraction vectorielle, la mesure de Dice s'écrit : $$Dice = \frac{2 \sum_i v_{i,A} \cdot v_{i,B}}{\sum_i v_{i,A} + \sum_i v_{i,B}} \quad \text{(II.10)}$$

Similarités liées aux modèles probabilistes	Similarités liées au principe d'ordonnancement probabiliste[3] : le modèle « binary independance model » ou le BM25. Similarités liées aux modèles de langue[4] : – le modèle de Ponte et Croft – le modèle de Hiemstra et al. – le modèle de Miller et al. – le ratio de vraisemblance – le modèle basée sur l'entropie croisée – le modèle de Kullback-Leibler – le modèle de rétro-pertinence – le modèle du canal bruité
Classification	Une classification consiste à partager une ensemble d'individus ou d'objets en sous-groupes appelés classes regroupant des individus ou des objets renfermant des informations communes et le plus dissemblables possible d'une classe à une autre. Il existe deux grands types de classification : – les méthodes non supervisées qui consistent à trouver les classes naturelles pour rassembler des données non étiquetées. Les classifications hiérarchiques (ascendantes ou descendantes), la méthode des nuées dynamiques [VFL+06], la méthode k-means sont des exemples de classifications non supervisées. – les méthodes supervisées qui consistent à apprendre une méthode pour prédire la classe d'un élément à partir d'éléments déjà classés. On trouve parmi ces méthodes : la méthode des k-plus proches voisins[IS07], les machines à vecteurs supports[Joa98], les classifieurs bayésiens naïfs [Seb02], les arbres de décisions [Seb02], les réseaux neuronaux [EGRCG+06].
Autre similarité	la distance intertextuelle proposée par Labbé et Labbé[LL03]

3. Pour plus d'informations sur ces modèles voir [CG11]
4. Pour plus d'informations sur ces modèles voir [BKN04b] et [CG11]

Similarités structurelles Pour les similarités structurelles, on peut citer les travaux de Champclaux[CDM10] :

Similarité basée sur SimRank	« *L'idée est donc de comparer des documents entre eux au travers des ressemblances entre les mots qu'ils contiennent ; les ressemblances entre les mots dépendant elles-mêmes des ressemblances entre les documents qui les contiennent* »[CDM07]. La similarité $S_d(d_i, d_j)$ entre deux documents d_i et d_j est définie comme suit : $$S_d(d_i, d_j) = \begin{cases} 1 & \text{si } d_i = d_j \\ \frac{C_1}{\lvert T_{d_i} \rvert \lvert T_{d_j} \rvert} \sum_{t_k \in d_i} \sum_{t_l \in d_j} S_t(t_k(d_i), t_l(d_j)) & \text{si } d_i \neq d_j \end{cases} \quad \text{(II.11)}$$ T_{d_i} est l'ensemble des termes du document d_i. $\lvert T_{d_i} \rvert$ est le nombre de termes appartenant au document d_i. $\lvert T_k(d_i) \rvert$ est le k$^{\text{ème}}$ terme du document d_i (le i$^{\text{ème}}$ document de la collection). C_1 est une constante de propagation. La similarité $S_t(t_i, t_j)$ entre deux termes t_i et t_j est définie comme suit : $$S_t(t_i, t_j) = \begin{cases} 1 & \text{si } t_i = t_j \\ \frac{C_2}{\lvert D_{t_i} \rvert \lvert D_{t_j} \rvert} \sum_{d_k \in D_{t_i}} \sum_{d_l \in D_{t_j}} S_d(d_k(t_i), d_l(t_j)) & \text{si } t_i \neq t_j \end{cases} \quad \text{(II.12)}$$ D_{t_i} est l'ensemble des documents contenant le terme t_i. $\lvert D_{t_i} \rvert$ est le nombre de documents contenant le terme t_i. $d_i(t_j)$ est le i$^{\text{ème}}$ document contenant le terme t_j (le j$^{\text{ème}}$ terme du vocabulaire). C_2 est une constante de propagation. « *Les formules traduisent le fait que la similarité de deux documents dépend de la similarité des termes qui les indexent ; et réciproquement la similarité de deux termes dépend de la similarité entre les documents dans lesquels ils apparaissent* »[CDM07].

2.5 Visualisation des résultats

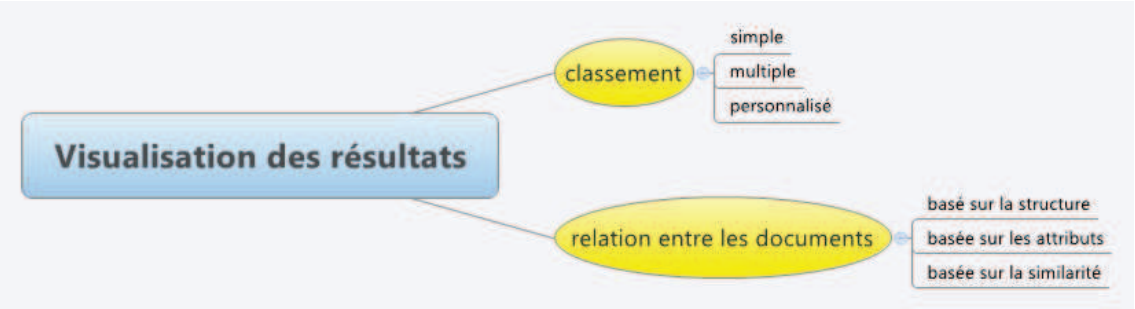


Figure II.9 – Visualisation des résultats

Cette fonctionnalité a pour objectif d'aider l'utilisateur à consulter et visualiser les documents qui répondent à ses besoins. Bonnel[Bon06] propose une classification des interfaces de visualisation présentée à la figure II.10.

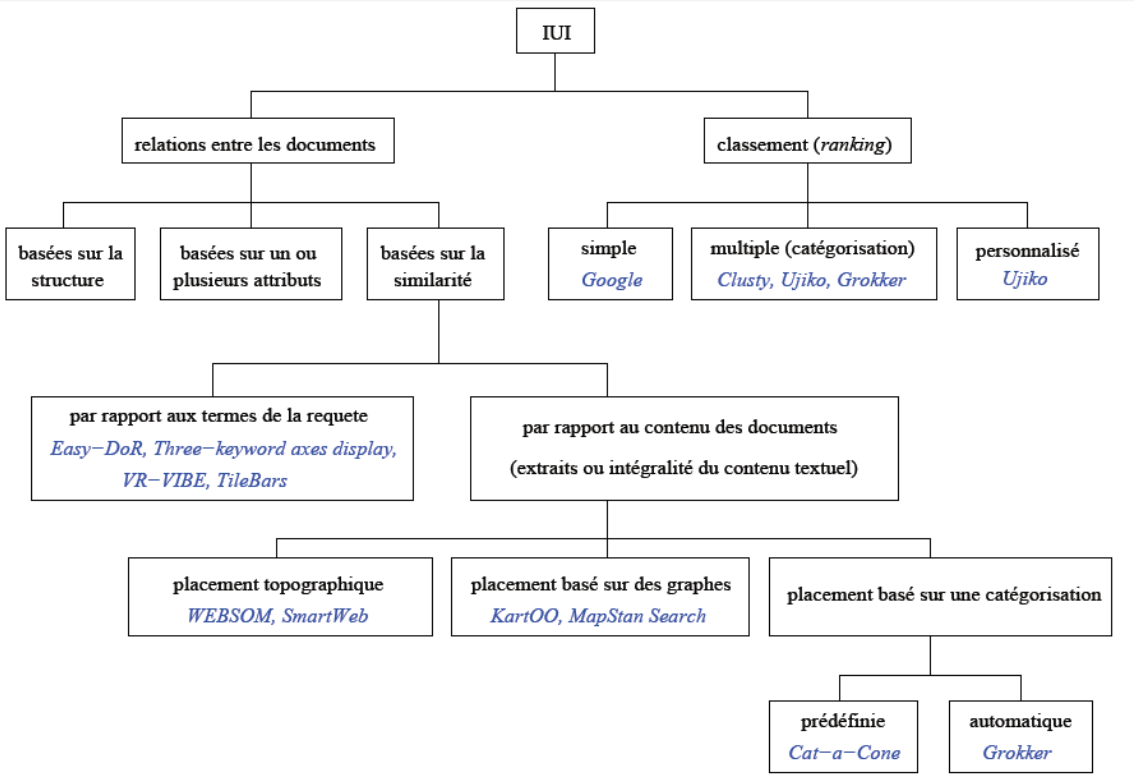


Figure II.10 – Classification des interfaces de visualisation basée sur l'organisation visuelle des résultats

Classement Les principaux moteurs de recherche (Google, Bing, Yahoo, ...) proposent une présentation des résultats sous forme de classement simple : une liste composée de liens hypertextes pour consulter les documents et un court descriptif (voir figure II.11). Ce mode de présentation des résultats est facile à mettre en œuvre

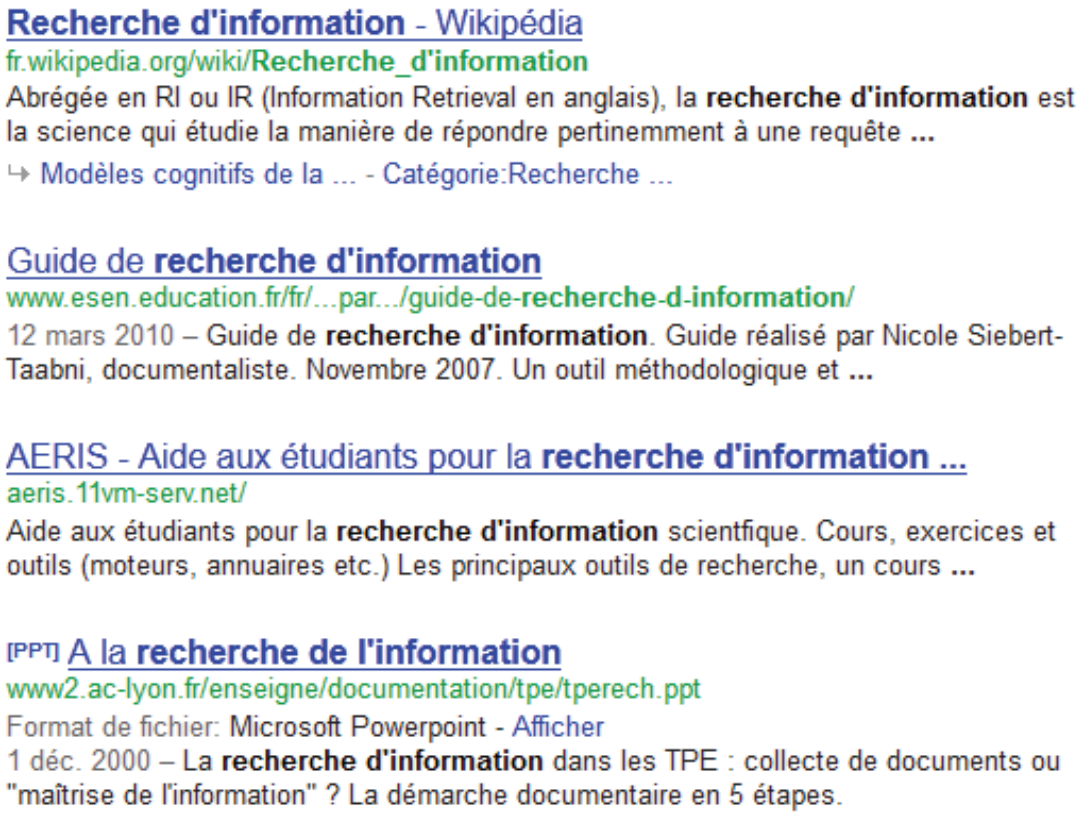


Figure II.11 – Liste de résultats issue de Google

mais il présente des limites :

- il n'est efficace que pour un nombre réduit de documents (inférieur à 20)[CLS00],
- l'absence de relations entre les résultats : l'utilisateur doit lire les documents pour établir les liens entre eux [Che02],
- « *la désorganisation thématique des résultats. Les résultats appartenant à différentes thématiques sont mélangés dans la liste de résultats* »[BCD08],
- la présentation en plusieurs pages des résultats ne permet pas à l'utilisateur

d'avoir une vision globale sur les résultats [BCD08].
D'autres moteurs de recherche tels que « Yippy » [5] proposent une classification des résultats en rubriques en plus de la simple liste (voir figure II.12). D'après Bonnel

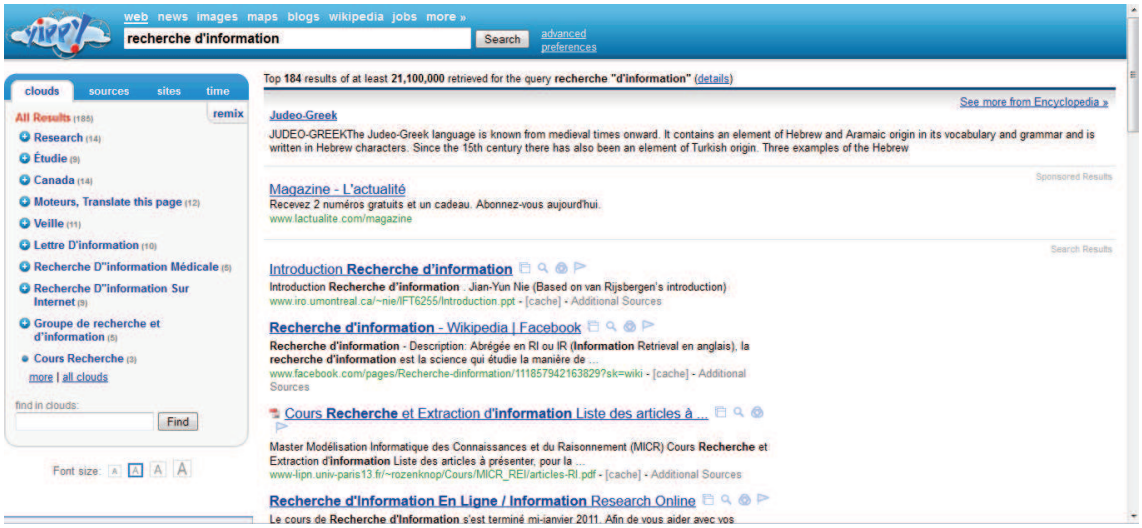


Figure II.12 – Liste de résultats issue de Yippi

et al.[BCD08], le moteur de recherche Ujiko de la société KartOO proposait un classement personnalisé. Malheureusement, la société KartOO a fermé ses moteurs de recherche en 2010 et nous n'avons pas pu tester ce type d'interface.

Relation entre les documents

Visualisations basées sur les attributs « *Il s'agit de techniques ayant pour objectif de visualiser les valeurs d'un ensemble d'attributs (c'est-à-dire les propriétés qualitatives et quantitatives des documents* »[BCD08]. Chevalier[Che02] présente *Cougar* qui permet de visualiser un ensemble de documents par rapport aux thèmes qu'ils abordent. Mothe et al. proposent « DocCube » : cet outil présente la répartition des documents en fonction d'entrées de hiérarchies de concepts. DocCube offre des visualisations globales d'information concernant la collection de documents couvrant le domaine choisi par l'utilisateur. La figure II.13 est composée d'un cube dont « *les axes [de ce] cube correspondent aux dimensions, c'est à dire aux hiérarchies de concepts. L'intersection des axes correspond au nombre de documents rattachés à*

5. http ://yippy.com/

cet axe. Cette information est représentée sous la forme d'une sphère dont la taille est proportionnelle au nombre de documents »[MCA01].

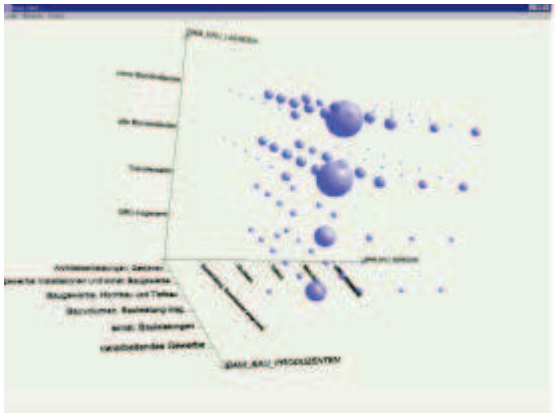


Figure II.13 – Liste de résultats issue de DocCube[MCA01]

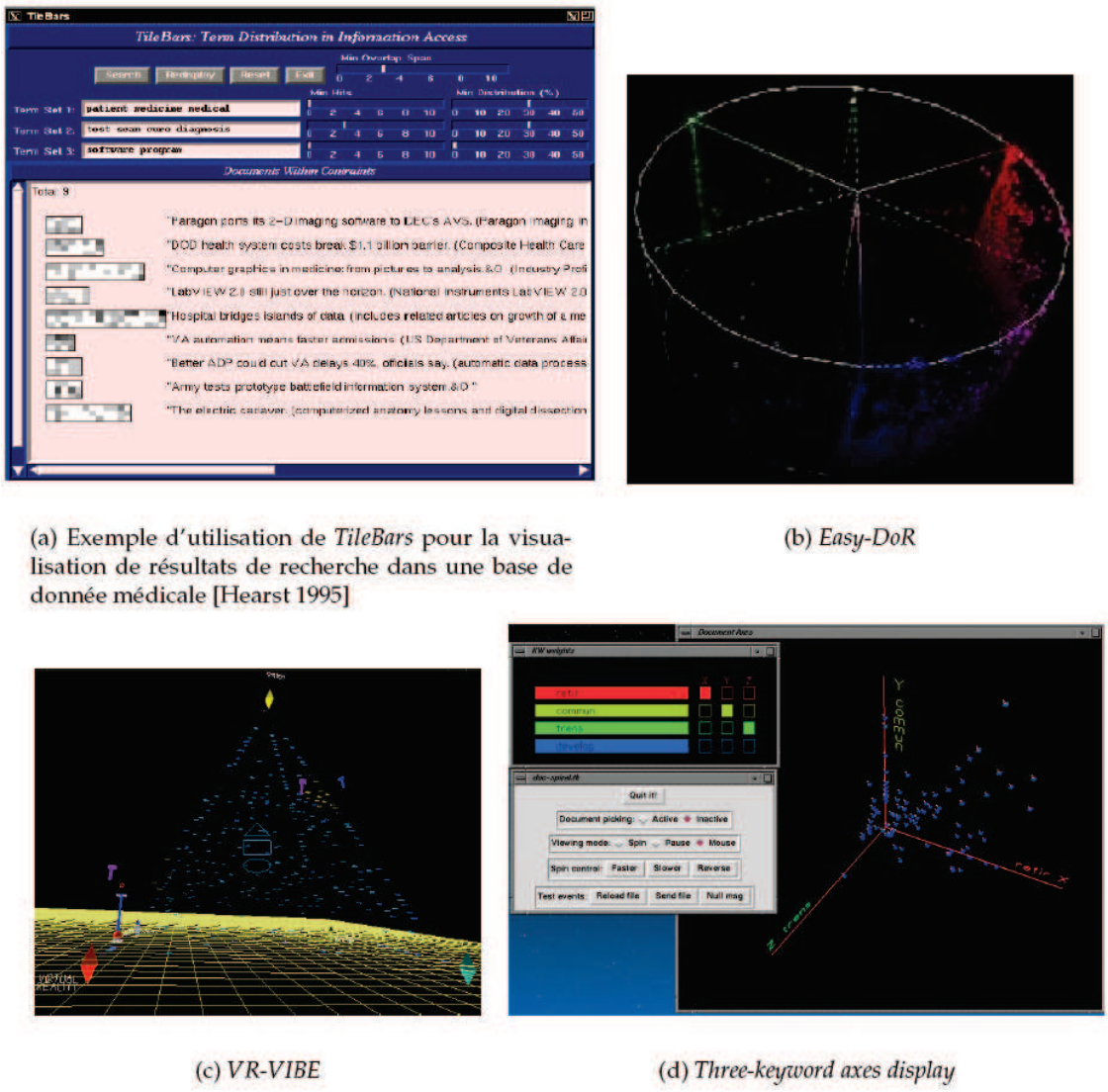
Ces représentations par concept permettent de visualiser les attributs d'un ensemble de documents mais elles ne permettent pas d'identifier rapidement les relations entre les documents.

Visualisations basées sur la similarité

Visualisations par rapport aux termes de la requête Ces visualisations présentent la répartition des différents mots-clés de la requête au sein des documents retrouvés. « *L'idée générale consiste à associer un mot-clé de la requête (ou une combinaison de mots-clés) à un axe ou une position de l'espace ; puis de positionner les documents dans cet espace en se basant, le plus généralement, sur la fréquence d'apparition du mot-clé dans le document. Ces techniques permettent donc d'identifier les relations entre les résultats et les mots-clés de la requête* »[Bon06]. Bonnel et al.[BCD08] présentent quatre outils (voir figure II.14) : Tiles-Bar, VR-VIBE, Three-keyword axes display, Easy-Dor. Ce type de visualisation permet d'identifier les relations entre les documents fournis et la requête. En revanche, les relations entre les documents ne sont pas identifiables.

Visualisation par rapport au contenu des documents Ces visualisations se découpent en trois groupes selon Bonnel[Bon06] :

– les approches cartographiques basées sur des graphes : ces approches disposent les documents pertinents sur une « carte » pour rendre le graphe résultat li-



(a) Exemple d'utilisation de *TileBars* pour la visualisation de résultats de recherche dans une base de donnée médicale [Hearst 1995]

(b) *Easy-DoR*

(c) *VR-VIBE*

(d) *Three-keyword axes display*

Figure II.14 – Liste de résultats issue d'outils de visualisation par rapport aux termes de la requête[Bon06]

sible. Les placements ne sont pas faits en tenant compte de la proximité sémantique entre les documents(voir figure II.15).

– les approches topographiques : ces approches, au contraire des approches cartographiques basées sur des graphes, tiennent compte de la distance sémantique entre les documents (voir figure II.16).
– les approches basées sur une catégorisation hiérarchique : chaque document est attribué à une catégorie voir sous-catégorie et l'importance de la catégorie

est mise en avant par la visualisation (voir figure II.17).

Figure II.15 – Liste de résultats issue d'outils de visualisation par rapport au contenu des documents (approches cartographiques) : MapStan

Les fonctionnalités présentées peuvent être intégrées dans des outils de recherche ou de navigation. Ces outils doivent être évalués : il est nécessaire de vérifier que les résultats offerts par ces outils sont pertinents.

Figure II.16 – Liste de résultats issue d'outils de visualisation par rapport au contenu des documents (approches topographiques) : City of News [SPD+97]

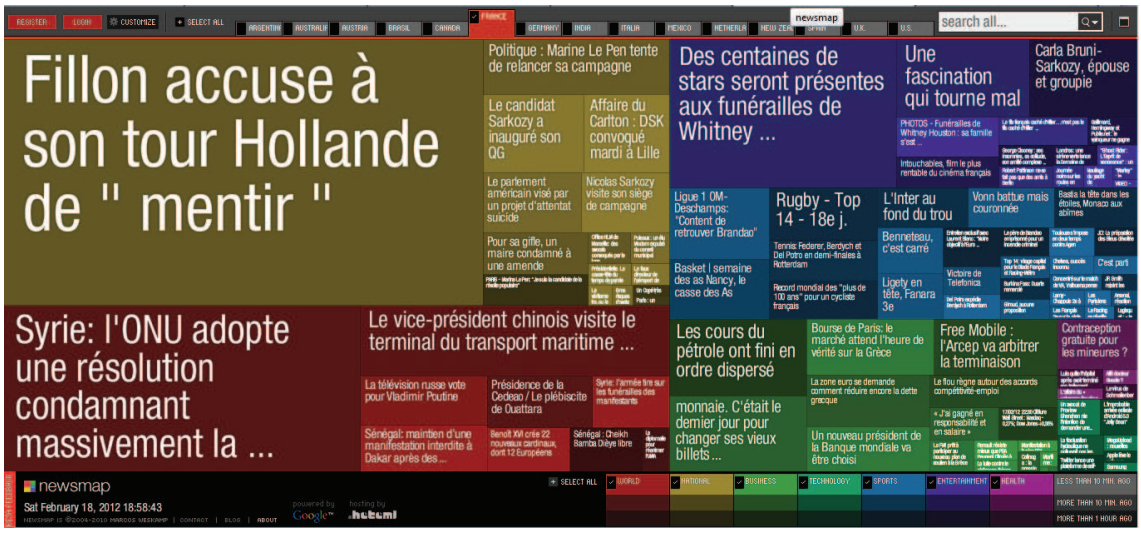


Figure II.17 – Liste de résultats issue d'outils de visualisation par rapport au contenu des documents (approches catégorisation hiérarchique) : NewsMap

3 Évaluation des outils

L'évaluation des outils de recherche d'information repose sur une notion centrale qui est la « pertinence ».

3.1 Pertinence

D'après Boughanem [BS08], concernant les outils de recherche d'informations, il existe deux types de pertinence :

- la pertinence système : elle est déterministe, objective et elle est définie à travers les modèles de recherche d'information.
- la pertinence utilisateur : elle est liée à la perception de l'utilisateur sur l'information renvoyée par le système. Elle est subjective, deux utilisateurs peuvent juger différemment un même document renvoyé par une même requête. Elle peut évoluer dans le temps d'une recherche.

L'objectif est de rapprocher le plus possible la pertinence système de la pertinence utilisateur.

Pincemin [Pin99] décrit plusieurs types de pertinence système :

- la pertinence binaire : pour un besoin d'information, un document est soit pertinent soit non pertinent,
- la pertinence n-aire : des degrés de pertinence sont ajoutés par rapport à la pertinence binaire : très pertinent, assez pertinent, peu pertinent, hors sujet, ... « *le système SPIRIT, dans ses versions récentes (Fluhr 1994) (SPIRIT-W3), propose une présentation intéressante des résultats d'une recherche sur une base documentaire. Ce qui l'apparente à une pertinence n-aire, c'est le fait que les documents sélectionnés soient regroupés en classes, et que ces classes sont elles-mêmes présentées par ordre de pertinence présumée décroissante. Le principe est le suivant : la requête consiste habituellement en quelques mots. La première classe est formée par les documents présentant tous les mots de la requête. La deuxième, par les documents présentant tous les mots sauf un (le moins "significatif" au sens d'indicateurs statistiques). Pour la troisième classe, les documents ont encore tous les mots sauf un, mais cette fois-ci le terme manquant était un peu plus important. Et ainsi de suite, jusqu'aux classes correspondant aux documents fournis en raison d'un seul terme* »[Pin99]

- la pertinence linéaire : un calcul de score est réalisé pour mesurer la relation entre une requête et un document, puis les documents sont ordonnés en fonction de ce score,
- la pertinence différentielle : elle « *fonctionne par regroupements et oppositions, et traduit ainsi les interrelations entre documents proposés en résultat de la recherche. Les regroupements traduisent les familles, qui peuvent être traitées collectivement (notamment pour mettre de côté des documents tous sélectionnés sur un aspect qui n'intéresse pas l'utilisateur). Les oppositions servent à contraster les documents les uns par rapport aux autres, pour mettre en valeur la singularité de chacun dans le contexte de cette requête et de ce fonds documentaire. C'est en effet une combinaison de jugements d'équivalence et de caractérisations spécifiques qui construit le résultat effectif de la recherche, et le choix motivé d'un ensemble de documents* »[Pin99].
- la pertinence polaire : « *La pertinence polaire correspond à une représentation spatiale, dans laquelle peuvent se dessiner plusieurs pôles d'attraction significatifs. Les documents se concentrent au niveau des différents pôles. Certains peuvent se positionner de façon intermédiaire, comme sous l'influence de différents pôles. Leur proximité relative à ces pôles traduit leur "attirance", potentiellement inégale* »[Pin99]

3.2 Outils d'évaluation

Les outils classiques d'évaluation en recherche d'informations vont mesurer l'écart entre la pertinence système et la pertinence utilisateur. Des experts établissent une liste de documents pertinents et une liste de documents non pertinents, soit en fonction d'une requête, soit en fonction d'une appartenance à une classe thématique. Puis ces listes vont être comparées aux listes de documents pertinents et non pertinents établies par l'application choisie.

Rappel Le rappel mesure la proportion de documents pertinents obtenus parmi tous les documents pertinents disponibles. Si le rappel vaut 1, cela signifie que tous les documents pertinents ont été fournis. Cette mesure permet de déterminer le silence i.e le nombre de documents pertinents non trouvés.

Pour un système de recherche par requête, le rappel se calcule alors :

$$Rappel = \frac{\text{Nombre de documents pertinents fournis}}{\text{Nombre de documents pertinents disponibles}} \tag{II.13}$$

Dans le cas d'un système présentant une classification, le rappel se calcule :

$$Rappel_i = \frac{documents\ correctement\ attribués\ à\ la\ classe\ i}{nombre\ de\ documents\ appartenant\ à\ la\ classe\ i} \tag{II.14}$$

$$Rappel = \frac{\sum_{i=1}^{n} rappel_i}{n} \tag{II.15}$$

Précision La précision mesure la proportion de documents pertinents fournis parmi tous les documents fournis. Elle mesure la capacité de l'outil à trouver des documents pertinents. Si la précision vaut 1, tous les documents fournis sont des documents pertinents. Cette mesure détermine le bruit i.e la proportion de documents non pertinents fournis par le système.
Pour un système de recherche par requête, la précision se calcule alors :

$$Précision = \frac{\text{Nombre de documents pertinents fournis}}{\text{Nombre de documents fournis}} \tag{II.16}$$

Dans le cas d'un système présentant une classification, la précision se calcule :

$$Précision_i = \frac{documents\ correctement\ attribués\ à\ la\ classe\ i}{nombre\ de\ documents\ attribués\ à\ la\ classe\ i} \tag{II.17}$$

$$Précision = \frac{\sum_{i=1}^{n} précision_i}{n} \tag{II.18}$$

Nous avons montré dans ce chapitre la grande diversité des outils et des techniques utilisées pour rechercher de l'information. Ces outils et techniques nécessitent plus ou moins de paramétrages (fixation d'un seuil par exemple), plus moins de temps

pour interpréter les résultats (par exemple les résultats présentés par DocCube sont difficilement lisibles) et plus ou moins d'assistance technique. L'objectif de la veille stratégique est d'obtenir un gain de temps pour la recherche d'informations voisines et par conséquent nous souhaitons proposer un outil :

- nécessitant le moins de paramétrage possible,
- nécessitant une assistance humaine faible,
- dont les résultats doivent pouvoir être facilement interprétables dans le cadre de la veille stratégique,
- utilisant le moins de ressources extérieures possibles : ontologie ou autres ressources sémantiques. Néanmoins ces outils sémantiques doivent pouvoir s'intégrer pour affiner les résultats.
- ne pouvant pas s'appuyer sur une structure prédéfinie (par exemple structure XML) car les informations de la veille stratégique proviennent de sources différentes et ont par conséquent des structures différentes.

Pour toutes ces raisons, nous nous attacherons, dans le chapitre suivant, à définir une mesure de proximité qui réponde au mieux aux attendes de l'équipe de veille stratégique. Nous la nommerons « mesure de voisinage ». Nous étudierons, dans les prochains chapitres, le comportement, les forces et les faiblesses de cette mesure.

Références bibliographiques

[Bal90] J.P. Balpe : *Hyperdocuments, hypertextes, hypermédias*. Eyrolles, 1990. 55

[Baz05] M. Baziz : *Indexation conceptuelle guidée par ontologie pour la recherche d'information*. Thèse de doctorat, [s.n.], [S.l.], 2005. 59, 65

[BCD08] N. Bonnel, M. Chevalier et B. Dousset : *Métaphores de visualisation des résultats de recherche d'information sur le web*. Lavoisier, 2008. 74, 75, 76

[Bes01] R. Besançon : *Intégration de connaissances syntaxiques et sémantiques dans les représentations vectorielles de textes*. Thèse de doctorat, Lausanne, 2001. 66

[Bis00] G. Bisson : La similarité : une notion symbolique/numérique. *Apprentissage symbolique-numérique (tome 2). Eds Moulet, Brito. Editions CEPADUES. pp. 169*, 201, 2000. 66

[BKL04] M.W. Bilotti, B. Katz et J. Lin : What works better for question answering : Stemming or morphological query expansion ? *In Proceedings of the Information Retrieval for Question Answering (IR4QA) Workshop at SIGIR 2004*, 2004. 58

[BKN04a] M. Boughanem, W. Kraaij et J.-N. Nie : *Modèles de langue pour la recherche d'information*, page 163–182. Lavoisier, 2004. 66

[BKN04b] M. Boughanem, W. Kraaij et J.Y. Nie : Modeles de langue pour la recherche d'information. *Les systemes de recherche d'informations*, pages 163–182, 2004. 71

[BLM03] J.-P. Barthélémy, X. Luong et S. Mellet : Prenons nos distances pour comparer des textes, les analyser et les représenter. *Corpus*, (2), 2003. 67

[Bon06] N. Bonnel : *Génération dynamique de présentations interactives en multimédia 3D, de données, pour les applications en ligne*. These, Université Rennes 1, décembre 2006. 73, 76, 77

[BP93] M. Bécue et R. Peiro : Les quasi-segments pour une classification automatique des réponses ouvertes. *Actes des 2ndes Journées Internationales d'analyse des données textuelles*, pages 310–325, 1993. 60

[Bru03] E. Brunet : Peut-on mesurer la distance entre deux textes ? *Corpus*, (2), 2003. 67

[BS08] M. Boughanem et J. Savoy : *Recherche d'information : état des lieux et perspectives*. Collection Recherche d'information et web, ISSN 1968-8008. Lavoisier, Paris : Hermès science publ., 2008. 51, 52, 59, 61, 62, 80

[BYRN99] R. Baeza-Yates et B. Ribeiro-Neto : *Modern information retrieval*. 1999. 49, 50, 54, 57

[CCT10] S.-S. Choi, S.-H. Cha et C.C. Tappert : A survey of binary similarity and distance measures. *Journal of Systemics, Cybernetics and Informatics*, 8(1):43–48, 2010. 69

[CDM07] Y. Champclaux, T. Dkaki et J. Mothe : Utilisation des similarités structurelles pour l'évaluation de la pertinence en recherche d'information. *In Colloque Veille Stratégique Scientifique et Technologique (VSST 2007), Marrakech (Maroc), 21/10/2007-25*, volume 10, page 2007, 2007. 72

[CDM10] Y. Champclaux, T. Dkaki et J. Mothe : An information retrieval models taxonomy based on an analogy between cognitive science and information retrieval. *In Actes colloque VSST'10*, 2010. 68, 72

[CG11] S. Clinchant et É. Gaussier : *Modèles probabilistes pour la recherche d'information*. Collection Recherche d'information et web, ISSN 1968-8008. Lavoisier, Paris : Hermès science publications, 2011. 53, 71

[Che02] M. Chevalier : *Interface adaptative pour l'aide à la recherche d'information sur le web*. These, Université Paul Sabatier - Toulouse III, Dec 2002. 74, 75

[CLS00] J. Cugini, S. Laskowski et M. Sebrechts : Design of 3-d visualization of search results- evolution and evaluation. *Visual data exploration and analysis VII*, pages 198–210, 2000. 74

[CV07] R. Cilibrasi et P. Vitanyi : The google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, 19(3):370–383, 2007. 66

[EGRCG+06] A. El Golli, F. Rossi, B. Conan-Guez, Y. Lechevallier *et al.* : Une adaptation des cartes auto-organisatrices pour des données décrites par un tableau de dissimilarités. *Revue de statistique appliquée*, 54(3):33–64, 2006. 71

[GY11] É. Gaussier et F. Yvon : *Modèles statistiques pour l'accès à l'information textuelle*. Collection Recherche d'information et web, ISSN 1968-8008. Lavoisier, Paris : Hermès science publications, 2011. 51

[Hei04] S. Heiden : Interface hypertextuelle à un espace de cooccurrences : implémentation dans weblex. *In* Anne Dister Gérard Purnelle, Cédrick Fairon, éditeur : *Le poids des mots*, volume 1, pages 577–588, Louvain-la-Neuve, Belgique, Mar 2004. Presses Universitaires de Louvain. 66

[Her06] N. Hernandez : *Ontologies de domaine pour la modélisation du contexte en recherche d'Information*. Thèse de doctorat, [s.n.], [S.l.], 2006. 59

[Iha04] M. Ihadjadene : *Les systèmes de recherche d'informations : modèles conceptuels*. Traité des sciences et techniques de l'information. Lavoisier, Paris : Hermès science publ., 2004. 52, 60

[IS07] F. Ibekwe-Sanjuan : *Fouille de texte*. Systèmes d'information et organisations documentaires. Hermès - Lavoisier, Mar 2007. 49, 50, 71

[IV98] N. Ide et J. Véronis : Introduction to the special issue on word sense disambiguation : the state of the art. *Comput. Linguist.*, 24:2–40, mars 1998. 65

[Joa98] T. Joachims : Text categorization with support vector machines : Learning with many relevant features. *Machine Learning : ECML-98*, pages 137–142, 1998. 71

[Joy09] J. Joyce : *Ulysses*. Echo Library, 2009. 62

[LC91] H. Le Crosnier : Une introduction à l'hypertexte. *BBF*, (4):280 – 294, 1991. 54

[Lem08] B. Lemaire : Limites de la lemmatisation pour l'extraction de significations. *In Actes des 9e Journées internationales d'Analyse Statistique des Données Textuelles (JADT'2008)*, pages 725–732, Lyon, France, 2008. 58

[LL03] C. Labbé et D. Labbé : La distance intertextuelle. *Corpus*, (2), 2003. 71

[Loi04] Y. Loiseau : *Recherche flexible d'information par filtrage flou qualitatif*. Thèse de doctorat, Université Paul Sabatier (Toulouse), 2004. 51, 52

[May05] D. Mayaffre : De la lexicométrie à la logométrie. *Astrolabe*, pages 1–11, 2005. 58

[MB09] S. Mellet et J.-P. Barthélemy : La topologie textuelle : légitimation d'une notion émergente. *Lexicometrica*, 7:http ://www.cavi.univ–paris3.fr/lexicometrica/numspeciaux/special9/mellet.pdf, 2009. 57

[MCA01] J. Mothe, C. Chrisment et J. Alaux : Visualisation globale de collections de documents sous forme d'hypercube. *Extraction des Connaissances et Apprentissage (ECA) journal*, 4:131–142, 2001. 76

[Mel01] S. Mellet : Lemmatisation et encodage grammatical : un luxe inutile ? *Lexicometrica*, (numéro spécial "Autour de la lemmatisation"), 2001. 58

[Mem00] D. Memmi : Le modèle vectoriel pour le traitement de documents. *Cahiers Leibniz*, (14), novembre 2000. 60, 63, 64, 65

[MM00] R. Mihalcea et D. Moldovan : Semantic indexing using wordnet senses. *In Proceedings of the ACL-2000 workshop on Recent advances in natural language processing and information retrieval : held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics - Volume 11*, RANLPIR '00, pages 35–45, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics. 65

[Moo48] C.N. Mooers : *Application of random codes to the gathering of statistical information*. Thesis, 1948. Thesis (M.S.) Massachusetts Institute of Technology. Dept. of Mathematics, 1948. 49

[MRS08] C.D. Manning, P. Raghavan et H. Schütze : *Introduction to information retrieval*. Cambridge University Press, Cambridge ; New York, 2008. 58, 61

[Nie08] J.-Y. Nie : Introduction à la RI : Indexation, 2008. 58

[Pin99] B. Pincemin : *Diffusion ciblée automatique d'informations : conception et mise en œuvre d'une linguistique textuelle pour la caractérisation des destinataires et des documents*. Thèse de doctorat, Paris IV, [S.l.], 1999. 56, 57, 58, 59, 62, 80, 81

[Por97] M. F. Porter : Readings in information retrieval. chapitre An algorithm for suffix stripping, pages 313–316. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997. 58

[PPL98] A. Pibarot, J. Picard et D. Labbé : Les syntagmes répétés dans l'analyse des commentaires libres. *S. Mellet (éd.) JADT*, pages 507–515, 1998. 60

[RJ08] S. Rosset et M. Jardino : Comparaison de documents : mesures de similarité et mesures de distance. june 2008. 68

[RK89] H. Ritter et T. Kohonen : Self-organizing semantic maps. *Biological Cybernetics*, 61:241–254, 1989. 10.1007/BF00203171. 64

[Sav93] J. Savoy : Stemming of french words based on grammatical categories. *Journal of the American Society for Information Science*, 44(1):1–9, 1993. 58

[Seb02] F. Sebastiani : Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47, 2002. 71

[SFW83] G. Salton, E.A. Fox et H. Wu : Extended boolean information retrieval. *Commun. ACM*, 26:1022–1036, November 1983. 52

[Sid02] S. Sidhom : *Plate-forme d'analyse morpho-syntaxique pour l'indexation automatique et la recherche d'information : de l'écrit vers la gestion des connaissances*. These, Université Claude Bernard - Lyon I, mars 2002. INSA DE LYON - EDIIS. 60

[SL83] A. Salem et P. Lafon : L'inventaire des segments répétés d'un texte. *Mots*, 6(1):161–177, 1983. 60

[SM83] G. Salton et M.J. McGill : *Introduction to modern information retrieval*. McGraw-Hill computer science series. McGraw-Hill, New York, 1983. 60, 61

[SPD+97] F. Sparacino, A. Pentland, G. Davenport, M. Hlavac et M. Obelnicki : City of news. *Ars Electronica Festival, Linz, Austria*, pages 8–13, 1997. 79

[ST09] G.M. Sacco et Y. Tzitzikas : *Dynamic Taxonomies and Faceted Search*. Springer, 2009. 49, 50

[TB89] P. Thoiron et H. Béjoint : Pour un index évolutif et cumulatif de cooccurrents en langue techno-scientifique sectorielle. *Meta*, 34(4): 661–671, 1989. 66

[Tom00] E.G. Toms : Understanding and facilitating the browsing of electronic text. *International Journal of Human-Computer Studies*, 52(3):423 – 452, 2000. 53, 54

[TR04] A. Tricot et J.-F. Rouet : Activités de navigation dans les systèmes d'information. *In Psychologie ergonomique : tendances actuelles*. Presses Universitaires de France - PUF, novembre 2004. 49

[Tri07] A.-P. Trinh : Classification de texte et estimation probabiliste par machine à vecteur support. *Actes de l'atelier DEFT07, CAp07*, 2007. 57

[VFL+06] A.M. Vercoustre, M. Fegas, Y. Lechevallier, T. Despeyroux *et al.* : Classification de documents xml à partir d'une représentation linéaire des arbres de ces documents. *et gestion des connaissances : EGC'2006*, 2006. 71

[Voo94] E.M. Voorhees : Query expansion using lexical-semantic relations. *In Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '94, pages 61–69, New York, NY, USA, 1994. Springer-Verlag New York, Inc. 65

Chapitre III

La mesure de voisinage

Dans ce chapitre, nous présentons la mesure de voisinage que nous avons élaborée à partir de la définition d'informations voisines. Nous décrivons la création de graphes à partir de la mesure ainsi que les propriétés de ces graphes. Nous comparons les résultats avec ceux obtenus à l'aide d'une mesure classique en recherche d'information le cosinus. Nous démontrons formellement les propriétés observées graphiquement à partir de textes aléatoires.

Sommaire

1 Fonctionnalités

Nous reprenons pour notre travail les fonctionnalités des systèmes de recherche d'informations décrites au chapitre II.

Abstraction En recherche d'information, l'approche classique repose sur la notion de « sac de mots » [BRN99][Tri07]. Nous avons choisi cette répresentation pour notre abstraction. Plus précisément, nous avons dans un premier temps lemmatisé les textes. La lemmatisation est une opération linguistique qui ramène les formes fléchies (conjuguées, plurielles, féminines) à une forme standard (infinitif, singulier, masculin). Elle a pour avantages de réduire le nombre de formes à traiter, d'affiner et stabiliser les traitements quantitatifs [Mel01]. Par exemple, le texte « *nous considérons un texte comme un tas de mots et nous avons choisi de représenter un texte comme un ensemble de mots lemmatisés* » sera représenté par l'ensemble des mots *{considérer, texte, tas, mot, choisir, représenter, ensemble, mot, lemmatiser}*.
Pour chaque base de textes, nous avons lemmatisé chaque texte à l'aide du logiciel Treetagger [1]. Nous avons choisi cet outil car :

- il est utilisé dans de nombreuses applications en recherche d'information ayant fait l'objet de publications,
- la lemmatisation est possible en plusieurs langues,
- la présence d'un algorithme sous la forme d'un arbre de décision lui permet de décider en contexte de la fonction grammaticale à attribuer au mot et d'en déduire le lemme le plus probable [2],
- il est gratuit.

Afin de ne pas alourdir les traitements et de pouvoir analyser des messages (tchat, SMS, ...), les mots mal orthographiés, imaginés ou inconnus du lemmatiseur sont conservés tels quels et ajoutés à la liste des mots lemmatisés. Nous noterons L_{T_i} la liste des mots lemmatisés, inconnus ou mal orthographiés du texte T_i.

1. TreeTagger est un outil qui permet d'annoter un texte avec des informations sur les parties du discours (genre de mots : noms, verbes, infinitifs et particules) et des informations de lemmatisation. Il a été développé par Helmut Schmid dans le cadre du projet « TC » dans le ICLUS (Institute for Computational Linguistics of the University of Stuttgart) http ://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html

2. avantage mis en avant par la société Onyme (blog.onyme.com/lemmatisation-et-racinisation-en-francais-flexion-lemme-et-racine-dun-mot/), société collaborant avec le centre de recherche en informatique de Lens (CRIL)

$	bis	depuis	extremis	-là	on	quatre-vingt-treize	soyons
£	bon	derechef	F	là-bas	-on	quatre-vingt-trois	stricto
a	C	derrière	facto	là-dedans	ont	quatre-vingt-un	suis
A	c’	des	fallait	là-dehors	onze	quatre-vingt-une	sur
à	ç’	dès	faire	là-derrière	or	que	sur-le-champ
afin	c.-à-d.	desdites	fais	là-dessous	ou	quel	surtout
ah	Ca	desdits	faisais	là-dessus	où	quelle	sus
ai	ça	désormais	faisait	là-devant	ouais	quelles	T
aie	çà	desquelles	faisaient	là-haut	oui	quelqu’	-t
aient	cahin-caha	desquels	faisons	laquelle	outre	quelque	t’
aies	car	dessous	fait	l’autre	P	quelquefois	ta
ailleurs	ce	dessus	faites	le	par	quelques	tacatac
ainsi	-ce	deux	faudrait	-le	parbleu	quelques-unes	tant
ait	céans	devant	faut	lequel	parce	quelques-uns	tantôt
alentour	ceci	devers	fi	les	par-ci	quelqu’un	tard
alias	cela	dg	flac	-les	par-delà	quelqu’une	te
allais	celle	die	fors	lès	par-derrière	quels	tel
allaient	celle-ci	différentes	fort	lesquelles	par-dessous	qui	telle
allait	celle-là	différents	forte	lesquels	par-dessus	quiconque	telles
allons	celles	dire	fortiori	leur	par-devant	quinze	tels
allez	celles-ci	dis	frais	-leur	parfois	quoi	ter
alors	celles-là	disent	fûmes	leurs	par-là	quoiqu’	tes

Tableau III.1 – Extrait de l’antidictionnaire français

Nous ne tenons pas compte de la fréquence d’apparition d’un mot dans un texte car nous considérons que tous les termes sont porteurs de sens. Ce choix est en lien avec la veille stratégique : un signal faible peut contenir un mot peu utilisé ou une association de mots peu fréquente [CFLBC10][LL11].

Filtrage Une fois la liste L_{T_i} des mots lemmatisés, inconnus ou mal orthographiés établie pour chaque texte T_i de la base, nous utilisons un antidictionnaire. Pour le français, nous nous sommes servis de celui mis à disposition par Véronis [3] pour filtrer les mots-outils tels que les auxiliaires avoir et être, les articles (le, la, les, ...), des adverbes, les chiffres, les symboles tels que $, la ponctuation...
Le tableau III.1 présente un extrait des mots utilisés dans l’antidictionnaire français.

Lissage Pour le lissage, nous utilisons un dictionnaire des synonymes ainsi qu’une mesure de la cooccurrence (présentée à la section 2.2) entre les mots dans notre base de textes. Nous avons choisi le dictionnaire des synonymes proposé par le site WebKeySoft [4] car :

3. professeur de linguistique et d’informatique, http ://sites.univ-provence.fr/veronis/donnees/index.html
4. http ://www.webkeysoft.com/articles/Dictionnaire-de-synonyme-francais-et-anglais.html

– ce dictionnaire est gratuit,
– il est facilement utilisable,
– il existe aussi une version anglaise.

Le dictionnaire de synonymes peut être évolutif : par exemple on peut ajouter des lignes si besoins. Par exemple, si notre base de textes concerne le thème CO_2, on peut ajouter une ligne « CO_2 ; gaz carbonique ; dioxyde de carbone ».

Mesure Pour répondre aux attendes de l'équipe de veille stratégique, nous avons construit une mesure de proximité entre textes, appelée mesure de voisinage, que nous présentons à la section 2. Cette mesure s'appuie sur la définition des informations voisines (voir chapitre I) : elle utilise les mots communs, les synonymes et la cooccurrence. Sa construction a été faite pour « coller » au mieux au problème concret de la veille stratégique.

Visualisation des résultats Nos résultats sont présentés sous forme de graphe permettant à l'utilisateur de naviguer dans les documents. La construction des graphes est exposée à la section 3. Pour tracer les graphes, nous avons utilisé le logiciel GraphViz [5].

2 Mesure de voisinage

2.1 Construction de la mesure de voisinage

Le calcul de notre mesure de voisinage entre deux textes est basé sur :

– les mots en commun : si un mot apparaît dans l'abstraction de chacun des textes
– les synonymes : si un mot pris dans l'abstraction du premier texte a au moins un synonyme dans l'abstraction du deuxième texte
– la cooccurrence : un mot d'une des listes de mots lemmatisés apparaît très fréquemment dans l'ensemble des textes de la base de données avec un mot de l'autre liste. Pour mesurer la cooccurrence entre deux mots, une distance est calculée à partir des textes où apparaissent les deux mots (seuls et en-

5. http ://www.graphviz.org/

semble). Par conséquent, plus deux mots sont cooccurrent, plus leur distance de cooccurrence est petite.

Prenons deux textes que nous notons T_i et T_j. Nous établissons la mesure de voisinage entre T_i et T_j de la manière suivante :

- nous établissons les listes L_{T_i} et L_{T_j} des mots lemmatisés, inconnus ou mal orthographiés de respectivement T_i et T_j,
- pour chaque mot de L_{T_i}, nous ajoutons :
  - 0 si le mot est aussi présent dans L_{T_j},
  - sinon *valeur_synonyme* si le mot a un synonyme dans L_{T_j},
  - sinon *valeur_cooccurrence_min* : on cherche le mot dans L_{T_j} pour lequel la *valeur_cooccurrence* est la plus petite,
- nous inversons les rôles de L_{T_i} et L_{T_j} et nous procédons à nouveau comme ci-dessus,
- pour terminer, nous ajoutons la somme sur les mots de L_{T_i} avec la somme sur les mots de L_{T_j} et nous divisons par 2.

Pour mesurer le cooccurrence, nous utilisons la distance définie par Cilibrasi et Vitanyi [CV06][CV07][CV04].

2.2 Calcul de la distance de Cilibrasi et Vitanyi

Définition

Cilibrasi et Vitanyi établissent une distance entre deux termes M_l et M_k basée sur le nombre de pages trouvées/indexées par Google contenant M_l, puis contenant M_k, contenant M_l et M_k ainsi que le nombre total de pages indexées par Google. Leur distance, notée DCV, est définie par :

$$DCV(M_l, M_k) = \frac{\max(\log(f(M_l), \log(f(M_k))) - \log(f(M_l, M_k))}{\log(M) - \min(\log(f(M_l), \log(f(M_{k2})))} \quad \text{(III.1)}$$

avec :

- $f(x)$ = nombre de pages contenant x
- $f(x, y)$ = nombre de pages contenant x et y
- M = nombre de pages indexées par Google.

Propriétés

1. Plus $DCV(x, y)$ est petite, plus les mots x et y sont cooccurrents.
2. Si $f(x), f(y) > 0$ et $f(x, y) = 0$ alors $DCV(x, y) = +\infty$
3. $DCV(x, y)$ n'est pas définie pour $f(x) = f(y) = 0$
4. $DCV(x, y) = \infty$ pour $f(x, y) = 0$
5. $DCV(x, y) > 0$ dans les autres cas.

Exemple

Nous reprenons ici l'exemple de Cilibrasi et Vitanyi dans [CV04] car Google ne fournit plus le nombre total de pages indexées.

- Nombre de pages contenant « rider » : 12 200 000.
- Nombre de pages contenant « horse » : 46 700 000
- Nombre de pages contenant « rider » et « horse » : 2 630 000
- Nombre de pages indexées par Google : 8 058 044 651

La distance DCV est alors :
$$DCV(rider, horse) = \frac{\max(\log(12200000), \log(46700000)) - \log(2630000)}{\log(8058044651) - \min(\log(12200000, 46700000)} \approx 0.443.$$

Dans leur article, Cilibrasi et Vitanyi ont initialement calculé la valeur à partir de Google mais nous calculons cette mesure sur notre base de textes. Nous reprenons leur formule en utilisant :

- $f(x)$ = nombre de textes de la base contenant x
- $f(x, y)$ = nombre de la base contenant x et y
- M = nombre de textes dans la base.

2.3 Calcul pour les synonymes

Dans notre calcul de voisinage, nous tenons compte des synonymes et nous pensons que la présence d'un synonyme dans un texte est plus importante que la présence d'un terme très cooccurrent. Il est donc nécessaire que la valeur pour un mot utilisé lors de la présence d'un synonyme soit plus grande que zéro (valeur pour les mots communs) et inférieure à la distance DCV la plus petite pour ce mot à n'importe quel mot du corpus. Afin d'obtenir une valeur qui respecte ces critères, nous avons

choisi de prendre la valeur la plus petite pour la distance *DCV* d'un mot avec tous les mots de la base de textes et de la diviser par deux. Nous avons fait ce choix pour pouvoir tracer l'évolution d'une information : par exemple une dépêche de l'AFP sera d'abord reprise telle quelle puis sera réécrite avec des synonymes.

2.4 Définition de la mesure de voisinage

Soit

- C un corpus,
- T_i et T_j deux textes
- L_{T_i} et L_{T_j} les listes des mots lemmatisés, inconnus ou mal orthographiés de respectivement T_i et T_j
- On note M_l un mot lemmatisé ou inconnu ou mal orthographié.

La mesure de voisinage est définie par :

$$MV(T_i, T_j) = \frac{\sum\limits_{M_l \in L_{T_i}} m(M_l, T_j) + \sum\limits_{M_k \in L_{T_j}} m(M_k, T_i)}{2} \tag{III.2}$$

avec :

$$m(M_l, T_j) = \begin{cases} 0 & \text{si } M_l \in T_j \\ \frac{\min\limits_{M_k \in C} DCV(M_l, M_k)}{2} & \text{si } T_j \text{ contient un synonyme de } M_l \quad valeur\_synonyme \\ \min\limits_{M_k \in L_{T_j}} DCV(M_l, M_k) & \text{sinon} \quad valeur\_cooccurrence\_min \end{cases} \tag{III.3}$$

Notre mesure est une mesure de proximité : les textes sont d'autant plus proches que la mesure est petite.

3 Représentations graphiques

3.1 Principe

Pour expliquer notre démarche concernant la représentation graphique liée à notre mesure, nous nous appuyons sur une base de textes : la base sur le cas « Automobile en Belgique ». Cette base a été constituée de 10 articles de journaux et 1 bulletin électronique (édité par l'ADIT).

Dans un premier temps, nous calculons la mesure de voisinage entre tous les textes présents dans notre base. Nous créons ainsi une matrice carrée de taille égale au nombre de textes. A la ligne i colonne j se trouve la mesure de voisinage MV entre le texte T_i et le texte T_j :

$$\begin{pmatrix} MV(T_1,T_1) & ... & MV(T_1,T_j) & ... & MV(T_1,T_{nombre\_de\_texte}) \\ ... & ... & MV(T_i,T_j) & ... & ... \\ MV(T_{nombre\_de\_texte},T_1) & ... & MV(T_{nombre\_de\_texte},T_j) & ... & MV(T_{nombre\_de\_texte},T_{nombre\_de\_texte}) \end{pmatrix}$$

Pour la base automobile, voici la matrice :

$$\begin{pmatrix} 0 & 31{,}63 & 22{,}87 & 23{,}20 & 49{,}67 & 43{,}94 & 25{,}11 & 34{,}89 & 37{,}71 & 41{,}21 & 32{,}79 \\ 31{,}63 & 0 & 29{,}32 & 29{,}43 & 58{,}13 & 46{,}07 & 45{,}89 & 44{,}09 & 38{,}55 & 36{,}91 & 38{,}76 \\ 22{,}87 & 29{,}32 & 0 & 21{,}79 & 50{,}05 & 40{,}45 & 24{,}74 & 19{,}63 & 23{,}40 & 36{,}10 & 31{,}12 \\ 23{,}20 & 29{,}43 & 21{,}79 & 0 & 44{,}96 & 37{,}53 & 23{,}87 & 21{,}23 & 33{,}40 & 20{,}51 & 31{,}16 \\ 49{,}67 & 58{,}13 & 50{,}05 & 44{,}96 & 0 & 54{,}52 & 46{,}01 & 44{,}49 & 52{,}69 & 47{,}09 & 52{,}76 \\ 43{,}94 & 46{,}07 & 40{,}45 & 37{,}53 & 54{,}52 & 0 & 41{,}75 & 38{,}27 & 47{,}14 & 38{,}22 & 44{,}30 \\ 25{,}11 & 45{,}89 & 24{,}74 & 23{,}87 & 46{,}01 & 41{,}75 & 0 & 21{,}19 & 20{,}98 & 18{,}92 & 32{,}20 \\ 34{,}89 & 44{,}09 & 19{,}63 & 21{,}23 & 44{,}49 & 38{,}27 & 21{,}19 & 0 & 15{,}12 & 17{,}16 & 28{,}84 \\ 37{,}71 & 38{,}55 & 23{,}40 & 33{,}40 & 52{,}69 & 47{,}14 & 20{,}98 & 15{,}12 & 0 & 10{,}55 & 53{,}30 \\ 41{,}21 & 36{,}91 & 36{,}10 & 20{,}51 & 47{,}09 & 38{,}22 & 18{,}92 & 17{,}16 & 10{,}55 & 0 & 50{,}51 \\ 32{,}79 & 38{,}76 & 31{,}12 & 31{,}16 & 52{,}76 & 44{,}30 & 32{,}20 & 28{,}84 & 53{,}30 & 50{,}51 & 0 \end{pmatrix}$$

Cette matrice contient toute l'information permettant de comparer tous les textes entre eux. Elle est très riche mais également plus difficile à exploiter. Pour simplifier, nous avons décidé de ne considérer que la plus petite mesure de voisinage entre un

texte et tous les autres textes de la base. Pour chaque ligne de la matrice des mesures de voisinage, nous avons cherché le minimum en enlevant la mesure de la diagonale (qui correspond à la mesure d'un texte à lui-même). Alors nous créons une nouvelle matrice avec des 1 pour les mesures minimales trouvées et 0 sinon : chaque ligne contiendra au moins un 1.

$$\begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

Une fois la matrice des minima calculée, nous créons le graphe associé : pour chaque texte, nous traçons une flèche vers le texte qui est le minimum. Par exemple, sur la base automobile, nous traçons une flèche du texte 1 vers le texte 3.
La figure III.1 présente le graphe ainsi obtenu.

3.2 Similarité cosinus

Pour comparer notre mesure à une similarité « classique » utilisée en recherche d'information, nous avons choisi la similarité cosinus.

3.2.1 Fonctionnalités

Nous appuyons ici également sur les fonctionnalités présentées au chapitre II.
L'abstraction pour la similarité cosinus repose sur le modèle vectoriel (chapitre II). Pour créer nos vecteurs, nous avons lemmatisé tous les textes de la base, filtré les

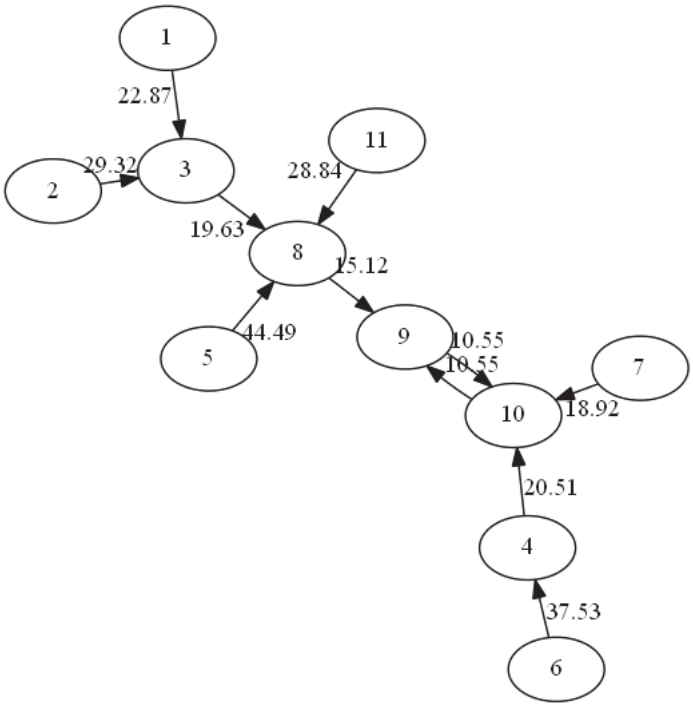


Figure III.1 – Graphe des minimums sur la matrice des mesures de voisinages

mots-outils à l'aide d'un antidictionnaire. De même que pour la mesure de voisinage, les mots mal orthographiés, imaginés ou inconnus du lemmatiseur sont conservés tels quels et ajoutés à la liste des mots lemmatisés. On constitue ainsi une liste de mots lemmatisés, mal orthographiés, imaginés et inconnus. Puis chaque mot de la liste constitue une composante du vecteur. Si la liste contient n mots, alors nos vecteurs textes seront de taille n. Pour calculer la valeur de chaque composante du vecteur d'un texte, nous avons utilisé deux méthodes :

- méthode booléenne : la composante est à 1 si le texte contient le mot associé à cette composante ; elle est à 0 sinon.
- méthode TF-IDF : on calcule le Term Frequency-Inverse Document Frequency(TF-IDF [BS08]) pour chacune des composantes :
  - TF terme frequency : $\mathrm{tf_{i,j}} = \frac{n_{i,j}}{\sum_k n_{k,j}}$
  - IDF inverse document frequency : $\mathrm{idf_i} = \log \frac{|D|}{|\{d_j : t_i \in d_j\}|}$

Ensuite nous calculons le cosinus pour deux vecteurs V_1 et V_2 de la manière suivante :

$$\cos(V_1, V_2) = \frac{V_1 \cdot V_2}{\|V_1\|\|V_2\|} \tag{III.4}$$

Le cosinus est une mesure de similarité : les textes sont d'autant plus proches que leur cosinus est grand (i.e proche de 1).

3.2.2 Représentation graphique

Nous procédons de manière analogue à la construction des graphes pour la mesure de voisinage. Nous construisons en premier la matrice des similarités cosinus entre textes.
Pour le cosinus sur des vecteurs booléens, nous obtenons sur notre exemple « Automobile en Belgique » :

$$\begin{pmatrix}
1 & 0{,}25 & 0{,}2 & 0{,}25 & 0{,}09 & 0{,}09 & 0{,}12 & 0{,}08 & 0{,}03 & 0{,}05 & 0{,}14 \\
0{,}25 & 1 & 0{,}27 & 0{,}24 & 0{,}11 & 0{,}16 & 0{,}08 & 0{,}1 & 0{,}11 & 0{,}17 & 0{,}16 \\
0{,}2 & 0{,}27 & 1 & 0{,}19 & 0{,}07 & 0{,}11 & 0{,}06 & 0{,}08 & 0{,}06 & 0{,}02 & 0{,}13 \\
0{,}25 & 0{,}24 & 0{,}19 & 1 & 0{,}12 & 0{,}14 & 0{,}1 & 0{,}1 & 0{,}04 & 0{,}15 & 0{,}11 \\
0{,}09 & 0{,}11 & 0{,}07 & 0{,}12 & 1 & 0{,}15 & 0{,}07 & 0{,}07 & 0{,}02 & 0{,}08 & 0{,}09 \\
0{,}09 & 0{,}16 & 0{,}11 & 0{,}14 & 0{,}15 & 1 & 0{,}08 & 0{,}07 & 0{,}03 & 0{,}15 & 0{,}13 \\
0{,}12 & 0{,}08 & 0{,}06 & 0{,}1 & 0{,}07 & 0{,}08 & 1 & 0{,}05 & 0{,}03 & 0{,}06 & 0{,}07 \\
0{,}08 & 0{,}1 & 0{,}08 & 0{,}1 & 0{,}07 & 0{,}07 & 0{,}05 & 1 & 0{,}09 & 0{,}06 & 0{,}09 \\
0{,}03 & 0{,}11 & 0{,}06 & 0{,}04 & 0{,}02 & 0{,}03 & 0{,}03 & 0{,}09 & 1 & 0{,}08 & 0{,}01 \\
0{,}05 & 0{,}17 & 0{,}02 & 0{,}15 & 0{,}08 & 0{,}15 & 0{,}06 & 0{,}06 & 0{,}08 & 1 & 0{,}07 \\
0{,}14 & 0{,}16 & 0{,}13 & 0{,}11 & 0{,}09 & 0{,}13 & 0{,}07 & 0{,}09 & 0{,}01 & 0{,}07 & 1
\end{pmatrix}$$

Pour le cosinus sur des vecteurs TF-IDF :

$$\begin{pmatrix}
1 & 0{,}1821 & 0{,}1819 & 0{,}0866 & 0{,}0094 & 0{,}0219 & 0{,}0267 & 0{,}0060 & 0{,}0088 & 0{,}0071 & 0{,}0410 \\
0{,}1821 & 1 & 0{,}2279 & 0{,}1037 & 0{,}0370 & 0{,}1128 & 0{,}0158 & 0{,}0095 & 0{,}0835 & 0{,}1599 & 0{,}0623 \\
0{,}1819 & 0{,}2279 & 1 & 0{,}0648 & 0{,}0128 & 0{,}0325 & 0{,}0126 & 0{,}0182 & 0{,}0263 & 0{,}0003 & 0{,}0372 \\
0{,}0866 & 0{,}1037 & 0{,}0648 & 1 & 0{,}0178 & 0{,}0453 & 0{,}0308 & 0{,}0448 & 0{,}0171 & 0{,}0504 & 0{,}0294 \\
0{,}0094 & 0{,}0370 & 0{,}0128 & 0{,}0178 & 1 & 0{,}0623 & 0{,}0240 & 0{,}0167 & 0{,}0026 & 0{,}0568 & 0{,}0366 \\
0{,}0219 & 0{,}1128 & 0{,}0325 & 0{,}0453 & 0{,}0623 & 1 & 0{,}0216 & 0{,}0635 & 0{,}0051 & 0{,}0432 & 0{,}0562 \\
0{,}0267 & 0{,}0158 & 0{,}0126 & 0{,}0308 & 0{,}0240 & 0{,}0216 & 1 & 0{,}0018 & 0{,}0011 & 0{,}0048 & 0{,}0175 \\
0{,}0060 & 0{,}0095 & 0{,}0182 & 0{,}0448 & 0{,}0167 & 0{,}0635 & 0{,}0018 & 1 & 0{,}0056 & 0{,}0022 & 0{,}0271 \\
0{,}0088 & 0{,}0835 & 0{,}0263 & 0{,}0171 & 0{,}0026 & 0{,}0051 & 0{,}0011 & 0{,}0056 & 1 & 0{,}0111 & 0{,}0025 \\
0{,}0071 & 0{,}1599 & 0{,}0003 & 0{,}0504 & 0{,}0568 & 0{,}0432 & 0{,}0048 & 0{,}0022 & 0{,}0111 & 1 & 0{,}0192 \\
0{,}0410 & 0{,}0623 & 0{,}0372 & 0{,}0294 & 0{,}0366 & 0{,}0562 & 0{,}0175 & 0{,}0271 & 0{,}0025 & 0{,}0192 & 1
\end{pmatrix}$$

Le cosinus est une mesure de similarité : par conséquent plus la valeur est grande (i.e plus elle est proche de 1) plus les textes sont similaires. Nous cherchons le maximum sur la ligne en enlevant la valeur sur la diagonale. Nous obtenons pour le cas booléen :

$$\begin{pmatrix}
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}$$

Pour le cas TF-IDF :

$$\begin{pmatrix}
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}$$

Une fois la matrice des maxima calculée, nous créons le graphe associé : pour chaque texte, nous traçons une flèche vers le texte qui est le maximum. Par exemple, sur la base automobile, nous traçons une flèche du texte 1 vers le texte 4 dans le cas booléen et de 1 vers 2 dans le cas TF-IDF.
Les graphiques obtenus sont présentés par les figures III.2 et III.3.

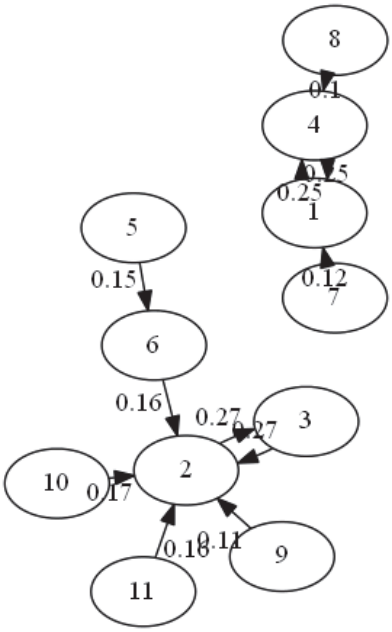


Figure III.2 – Graphe des maxima pour la matrice de similarité cosinus sur des vecteurs booléens

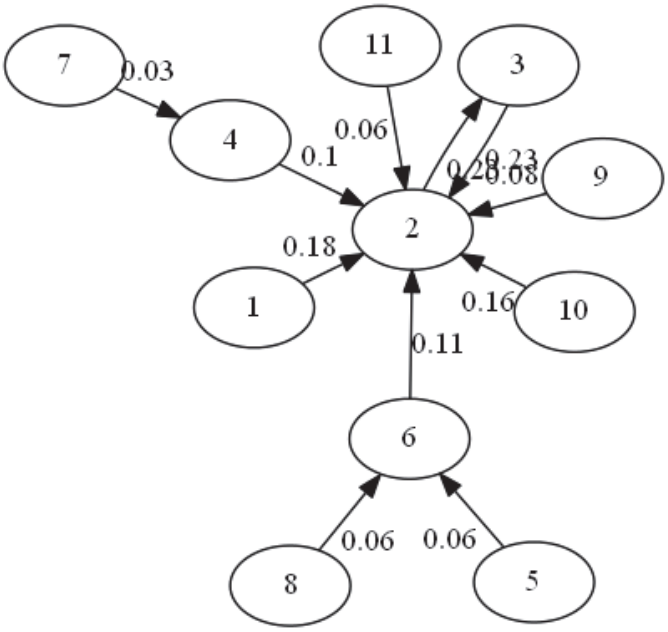


Figure III.3 – Graphe des maxima pour la matrice de similarité cosinus sur des vecteurs TF-IDF

3.3 Présentation des bases de textes utilisées

Pour tester les différentes mesures, nous avons constitué trois bases de textes :

Textes générés aléatoirement (TA) Pour constituer cette base, nous avons créé 300 textes aléatoires de la manière suivante :

- génération aléatoire de la taille T du texte (entre 300 et 1000 mots)
- tirage aléatoire avec remise de T mots dans un dictionnaire

Le dictionnaire utilisé contient 88349 « mots » différents : ce dictionnaire contient en plus du verbe à l'infinitif, ses conjugaisons.

Textes variés (TV) Cette base a été constituée à l'aide de textes choisis au hasard sur différents thèmes et de différents styles : des poèmes, des textes de loi, des recettes de cuisine, des nouvelles, des discours politiques, des petites annonces, des exercices de style de Raymond Queneau et des critiques de livres et de films.

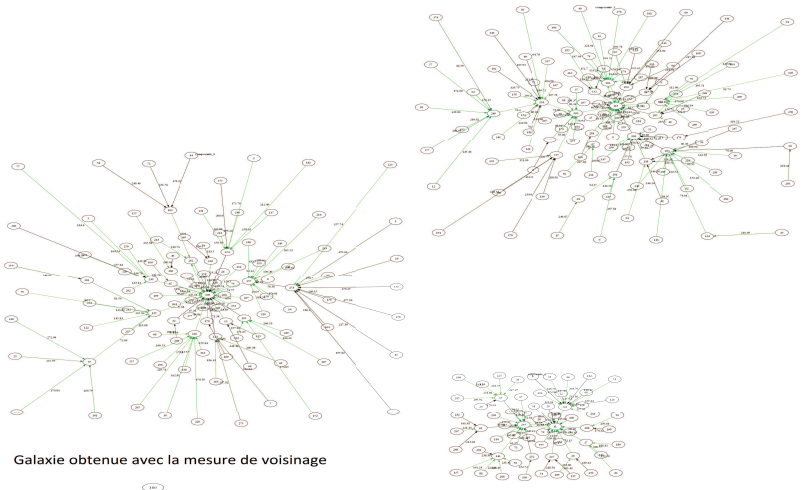
Textes sur une thématique : le cas CO_2 (CO_2) Cette base a été constituée dans une premier temps avec des articles obtenus à l'aide de Factiva [6] contenant les termes « chimie verte » ou « chimie durable » puis « valorisation du CO_2 ». Cette base contient 299 textes contenant entre 30 et 1052 racines différentes. Cet exemple sera détaillé plus largement dans le chapitre V.

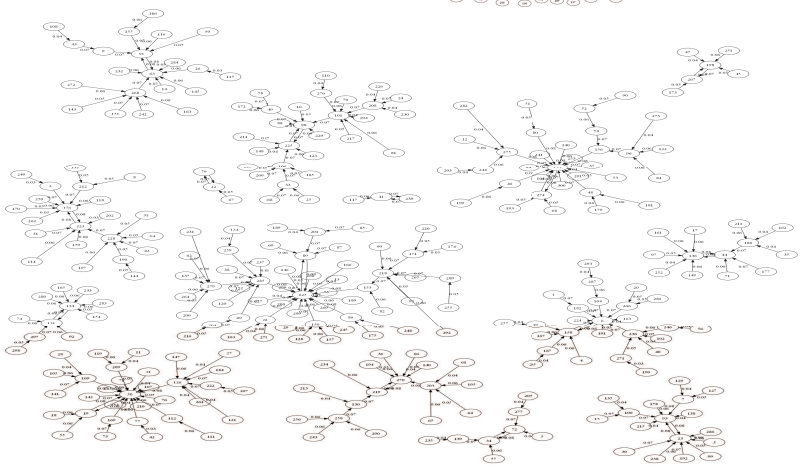
3.4 Graphes obtenus

Les figures III.4, III.5 et III.6 présentent les graphes obtenus [7] sur les différentes bases de textes à partir de la mesure de voisinage et du cosinus.

6. « Factiva (société Dow Jones & Company) agrège des contenus provenant à la fois de sources sous licence et gratuites, et apporte aux entreprises des fonctionnalités de recherche, d'alerte, de diffusion et de gestion de l'information. Les produits Factiva donnent accès à plus de 28 500 sources (comme des journaux, magazines retranscriptions radio et télévision, photos, etc..) » (source : Wikipédia)

7. ces graphiques sont donnés chacun en pleine page en annexes





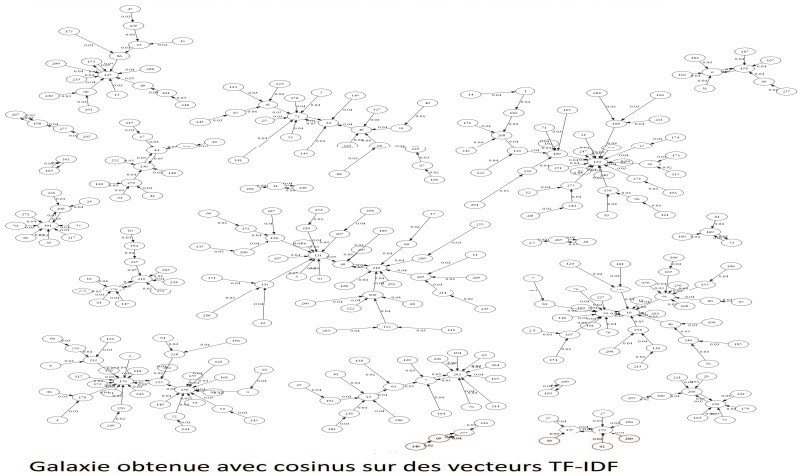
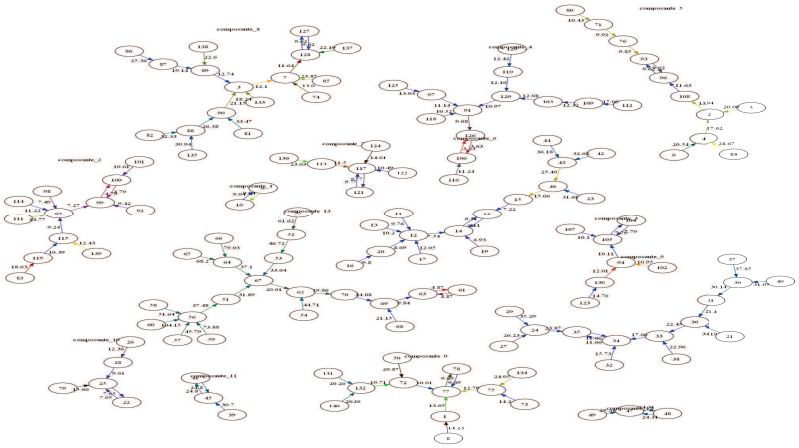
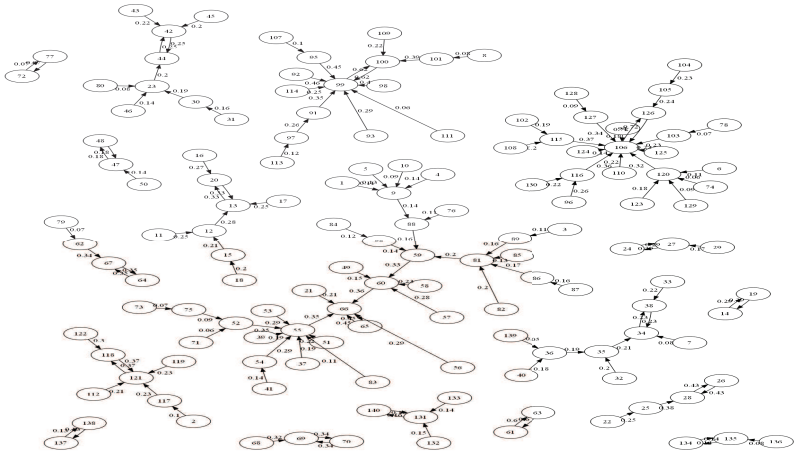


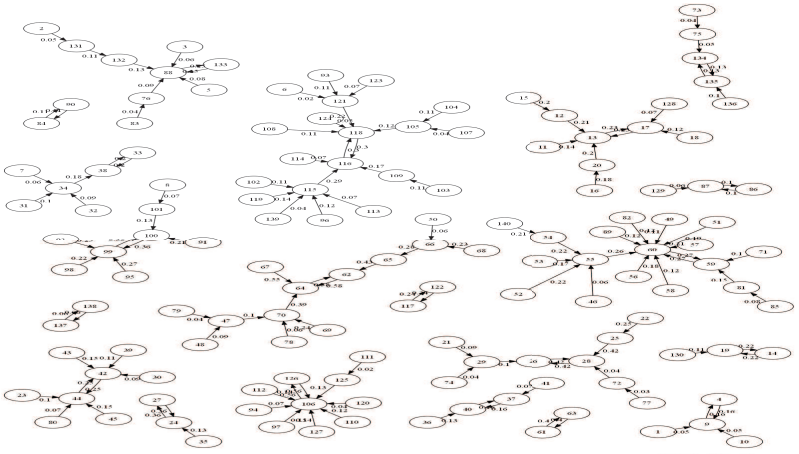
Figure III.4 – Textes aléatoires



Galaxie obtenue avec la mesure de voisinage

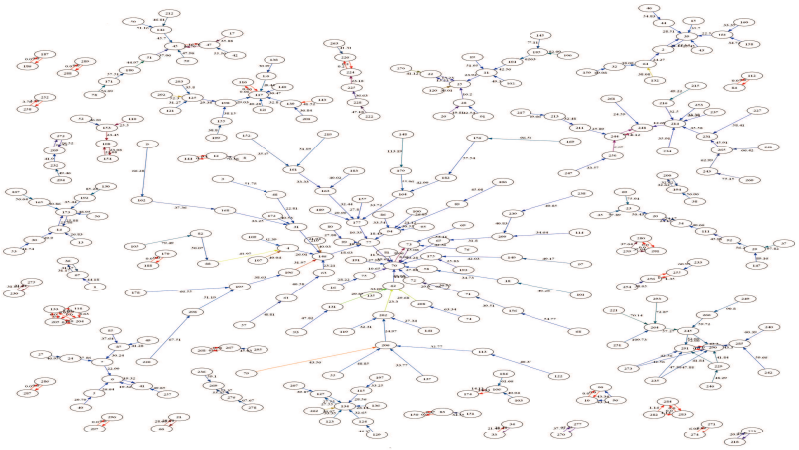


Galaxie obtenue avec cosinus sur des vecteurs booléens

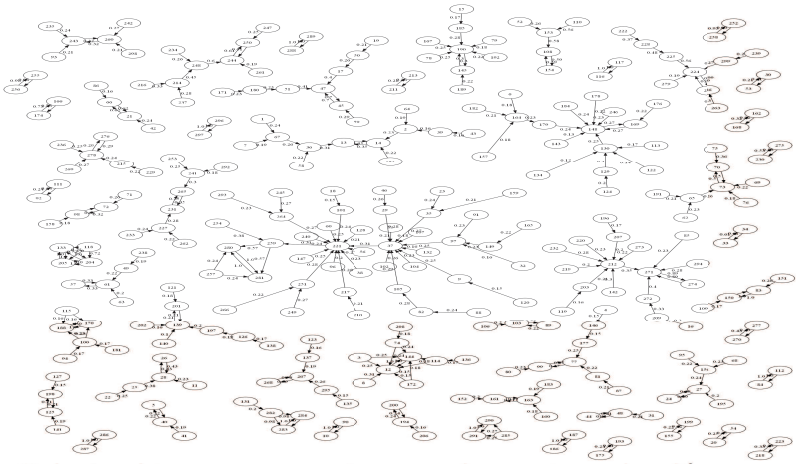


Galaxie obtenue avec cosinus sur des vecteurs TF-IDF

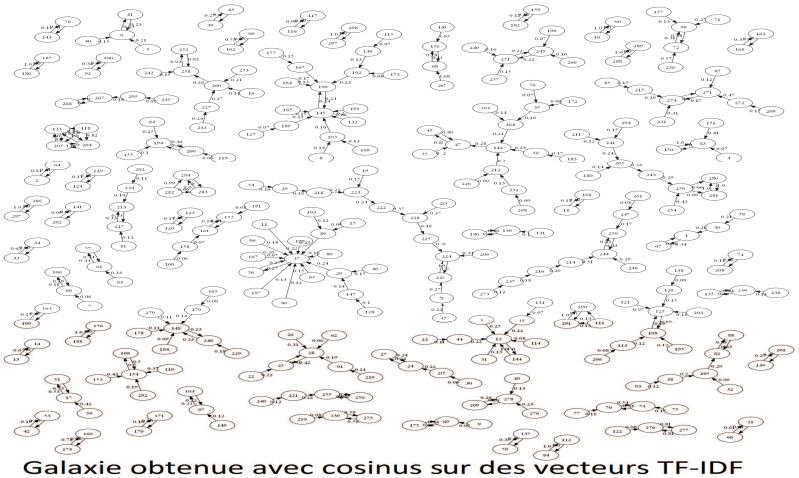
Figure III.5 – Textes variés



Galaxie obtenue avec la mesure de voisinage



Galaxie obtenue avec cosinus sur des vecteurs booléens



Galaxie obtenue avec cosinus sur des vecteurs TF-IDF

Figure III.6 – Textes CO2

3.5 Observations graphiques et statistiques

	Cosinus booléen			Cosinus TFIDF			Mesure de voisinage		
Base de données	TA	TV	CO_2	TA	TV	CO_2	TA	TV	
Nombre de composantes connexes	15	18	53	20	20	67	3	14	39
Moyenne des mesures similarités dans le graphe	0,06	0,24	0,42	0.04	0,17	0,4	155,24	19,64	31,59
Chemins descendants									
Longueur maximale	4	5	6	6	3	7	4	7	6
Moyenne des longueurs	1,73	2,11	2,12	1,82	1,55	1,68	2,04	3,05	2,48
Nombre de textes ayant un min/max dont le nombre de mots est inférieur	23 (7,7%)	57 (40,7%)	102 (34,1%)	29 (9,6%)	49 (35%)	112 (37,5%)	297 (99%)	111 (79,2%)	227 (75,9%)
Arcs entrants									
max	17	10	12	14	9	11	31	4	10

Tableau III.2 – Caractéristiques des graphes des différentes bases de données étudiées

Grâce aux statistiques données dans le tableau III.2 et à l'observation des graphes , nous remarquons que :

- lorsque la base de textes porte sur un thème plus restreint (ici base CO_2), le nombre de composantes connexes (ligne « Nombre de composantes connexes ») est plus important et les mesures sont plus petites (pour la mesure de voisinage) respectivement plus grandes (pour le cosinus). La partition obtenue est donc d'autant plus fine que le thème est commun. Ceci s'explique par de meilleurs résultats sur la présence de mots communs, de synonymes et sur la distance DCV qui mesure la cooccurrence,
- les chemins descendant vers les cycles (ou nucléus) sont en moyenne plus longs pour la mesure de voisinage,
- le nombre de composantes connexes est moins important pour la mesure de voisinage que pour le cosinus,
- Dans les graphes liés à la mesure cosinus, les textes pointent, majoritairement) sur des textes qui comptent plus de mots qu'eux (voir la case « Nombre de textes ayant un min/max dont le nombre de mots est inférieur »). La mesure cosinus propose une « synthèse » des documents : on va du particulier vers le général (de textes courts vers des textes plus longs),
- au contraire, nous avons cherché à construire une mesure qui place les textes ayant le plus de mots aux extrémités des constellations locales. Dans les che-

mins descendants, le nombre de mots composant les textes diminue en se rapprochant des cycles pour la mesure de voisinage. La mesure de voisinage propose une « analyse » des documents : on va du général vers le particulier (de textes longs vers des textes plus courts),
- grâce à la mesure de voisinage et sa propriété d'« analyse », les nucléus permettent un étiquetage thématique des composantes. Ceci intéresse vivement l'équipe de veille stratégique.

4 Justification des propriétés graphiques

Nous avons observé sur les graphes que :
- sur ceux tracés à partir de la mesure de voisinage, le long d'un bras (en direction du nucléus) :
  - la mesure de voisinage était de plus en plus petite,
  - le nombre de mots des textes diminuait,
- sur ceux tracés à partir du cosinus, le long d'un bras (en direction du nucléus) :
  - la mesure de voisinage était de plus en plus grande,
  - le nombre de mots des textes augmentait,

Nous allons démontré mathématiquement ces propriétés.

4.1 Graphes liés à la mesure de voisinage

4.1.1 Décroissance de la mesure de voisinage

Proposition 4.1. *soit T_i, T_j et T_k trois textes tels T_j est le texte minimum pour T_i et tels que T_k est le texte minimum pour T_j, alors la mesure de voisinage entre T_i et T_j est plus grande que celle entre T_j et T_k.*

Démonstration. Par l'absurde, on note :
- T_i, T_j et T_k trois textes
- $MV(x, y)$ la mesure de voisinage entre x et y

On suppose que le texte T_j est le texte le plus proche de T_i et T_k le plus proche de T_j. On a : $T_i \xrightarrow{\text{MV(Ti,Tj)}} T_j \xrightarrow{\text{MV(Tj,Tk)}} T_k$
Si on fait l'hypothèse que $MV(T_i, T_j) < MV(T_j, T_k)$, alors le texte T_i est le texte le plus proche de T_j ce qui contredit notre hypothèse de départ. □

4.1.2 Propriété du nucléus

Corollaire 4.2. *Le nucléus réalise le minimum de la constellation locale.*

Démonstration. Une constellation locale est composée d'un nombre fini de textes et donc d'un nombre fini de valeurs (mesures de voisinage). D'après la proposition 4.4, dans chaque bras de la constellation locale il y décroissance de la mesure. Comme nous avons un nombre fini de valeurs, il existe au moins une valeur minimale réalisée pour un texte que nous notons T_i. Cette mesure minimale est réalisée entre le texte T_i et un autre texte que nous notons T_j. Par définition, notre mesure de voisinage est symétrique : $MV(T_i, T_j) = MV(T_j, T_i)$. Par conséquent, le nucléus composé de T_i et T_j réalise le minimum de la constellation locale . □

4.1.3 Décroissance du nombre de mots

Théorème 4.3. *Pour des textes aléatoires la mesure a la propriété suivante si $T_i \rightarrow T_j$ dans une constellation et si $\#T_i < \#D/2$ pour le cas sans synonyme (resp $\#T_i < \#D/(c_s+2)$ pour le cas avec synonymes) où $\#D$ représente le nombre de mots dans le dictionnaire et c_s le nombre moyen de synonymes, alors pour minimiser la mesure entre les textes T_i et T_j, il faut que le nombre de mots dans T_i soit plus grand que le nombre de mots dans T_j.*

Démonstration. Soit :

- D un dictionnaire
- T_i et T_j deux textes
- M_l un mot de T_i
- $S(M_l)$ l'ensemble des synonymes de M_l

On note :

- $\#D$ le cardinal de D
- $\#T_i$ et $\#T_j$ les cardinaux de T_i et T_j

On suppose que pour deux mots quelconques, on trouve toujours la même distance de Cilibrasi et Vitanyi entre ces deux mots (supposition correcte pour les textes aléatoires). On note c_{DCV} cette distance.

En utilisant (III.2), on a :

$$\begin{aligned} E(MV(T_i,T_j)) &= E\left(\frac{\sum\limits_{M_l\in T_i} m(M_l,T_j)+\sum\limits_{M_k\in T_j} m(M_k,T_i)}{2}\right) \\ &= \frac{1}{2}\left(E(\sum_{M_l\in T_i} m(M_l,T_j)+\sum_{M_k\in T_j} m(M_k,T_i)\right) \\ &= \frac{1}{2}\left(\sum_{M_l\in T_i} E(m(M_l,T_j))+\sum_{M_k\in T_j} E(m(M_k,T_i))\right) \end{aligned} \tag{III.5}$$

Cas sans synonyme

Soit $M_l \in T_i$

On veut évaluer la probabilité que le mot M_l apparaisse dans le texte T_j sachant que le texte T_j est de taille fixée à k.

On va commencer par les petites valeurs :

- pour $k=1$: $P(M_l|\#T_j=1)=\dfrac{1}{\#D}$ et donc $P(\overline{M_l}|\#T_j=1)=1-\dfrac{1}{\#D}$
- pour $k=2$: $P(M_l|\#T_j=2)=\dfrac{2}{\#D}$ et donc $P(\overline{M_l}|\#T_j=2)=1-\dfrac{2}{\#D}$
- ...

Pour k fixé $<\#D$, on a :

	$M_l \in T_j$	$M_l \notin T_j$
$m(M_l,T_j)$	0	c_{DCV}
Probabilité	$\dfrac{\#T_j}{\#D}$	$1-\dfrac{\#T_j}{\#D}$

En utilisant (III.3), on peut écrire :

$$\begin{aligned} E(m(M_l,T_j) &= E(\min_{M_k\in T_j} DCV(M_l,M_k))) \\ &= 0\times\frac{\#T_j}{\#D}+c_{DCV}\times(1-\frac{\#T_j}{\#D}) \\ &= c_{DCV}\times(1-\frac{\#T_j}{\#D}) \end{aligned} \tag{III.6}$$

Donc (III.5) devient :

$$\begin{aligned} E(MV(T_i,T_j)) &= \frac{1}{2}\left(\#T_i \times c_{DCV} \times (1-\frac{\#T_j}{\#D}) + \#T_j \times c_{DCV} \times (1-\frac{\#T_i}{\#D})\right) \\ &= \frac{c_{DCV}}{2\#D}\left(\#D\#T_i + \#D\#T_j - 2\#T_i\#T_j\right) \end{aligned} \tag{III.7}$$

Si on fixe la taille de T_i, on peut écrire :

$$E(MV(T_i,T_j)) = \frac{c_{DCV}}{2\#D}\left(\#D\#T_i + \#T_j(\#D - 2\#T_i)\right) \tag{III.8}$$

Si $\#T_i < \frac{\#D}{2}$, alors pour minimiser la mesure entre deux textes, il faut prendre $\#T_j$ le plus petit possible.
Si $\#T_i = \frac{\#D}{2}$, alors le résultat est constant : il ne dépend pas de T_j.
Si $\#T_i > \frac{\#D}{2}$, alors pour minimiser la mesure entre deux textes, il faut prendre $\#T_j$ le plus grand possible.

En conclusion, si un texte T_i ne contient pas plus de la moitié des mots du dictionnaire, le texte T_j le plus proche aura moins de mots. Dans le cas d'un texte en français, il est peu probable qu'un texte contienne plus de la moitié des mots de la langue française.

Cas avec les synonymes
Soit $M_l \in T_i$
On suppose que le nombre de synonymes pour un mot $M_l \in T_i$ est constant : on note c_s cette constante. On va commencer par les petites valeurs :

- pour $k = 1$: $P(M_l \textit{ ait un synonyme dans } T_j | \#T_j = 1)$
  $= \textit{probabilité d'avoir un synonyme dans le dictionnaire} * \textit{taille de } T_j = \frac{c_s}{\#D} * 1$
- pour $k = 2$: $P(M_l \textit{ ait un synonyme dans } T_j | \#T_j = 2) = \frac{2 * c_s}{\#D}$
- ...

Pour k fixé $< \#D$, on a :

	$M_l \in T_j$	$M_l \notin T_j$	
		$S(M_l) \cap T_j \neq \varnothing$	$S(M_l) \cap T_j = \varnothing$
$m(M_l, T_j)$	0	$\frac{c_{DCV}}{2}$	c_{DCV}
Probabilité	$\frac{\#T_j}{\#D}$	$\frac{c_s \#T_j}{\#D}$	$1 - \frac{\#T_j}{\#D} - \frac{c_s \#T_j}{\#D}$

En utilisant (III.3), on peut écrire :

$$\begin{aligned} E(m(M_l, T_j)) &= E(\min_{M_k \in T_j} DCV(M_l, M_k)) \\ &= 0 \times \frac{\#T_j}{\#D} + \frac{c_{DCV}}{2} \times \frac{c_s \#T_j}{\#D} + c_{DCV} \times \left(1 - \frac{\#T_j}{\#D} - \frac{c_s \#T_j}{\#D}\right) \\ &= \frac{c_{DCV}}{\#D}\Big(\#D - \#T_j(\frac{c_s}{2} + 2)\Big) \end{aligned} \tag{III.9}$$

Donc (III.5) devient :

$$\begin{aligned} E(MV(T_i, T_j)) &= \frac{c_{DCV}}{2\#D}\left(\#T_i\Big(\#D - \#T_j(\frac{c_s}{2} + 1)\Big) + \#T_j\Big(\#D - \#T_i(\frac{c_s}{2} + 1)\Big)\right) \\ &= \frac{c_{DCV}}{2\#D}\left(\#T_i \#D + \#T_j\Big(\#D - \#T_i(c_s + 2)\Big)\right) \end{aligned} \tag{III.10}$$

Si $\#T_i < \dfrac{\#D}{c_s + 2}$, alors pour minimiser la mesure entre deux textes, il faut prendre $\#T_j$ le plus petit possible.
Si $\#T_i = \dfrac{\#D}{c_s + 2}$, alors le résultat est constant.
Si $\#T_i > \dfrac{\#D}{c_s + 2}$, alors pour minimiser la mesure entre deux textes, il faut prendre $\#T_j$ le plus grand possible.

□

4.2 Graphes liés au cosinus

4.2.1 Croissance de la mesure de voisinage

Proposition 4.4. *soit T_i, T_j et T_k trois textes tels T_j est le texte maximum pour T_i et tels que T_k est le texte maximum pour T_j, alors le cosinus entre T_i et T_j est plus petit que celui entre T_j et T_k.*

Démonstration. analogue à la démonstration pour la mesure de voisinage □

4.2.2 Propriété du nucléus

Corollaire 4.5. *Le nucléus réalise le maximum de la constellation locale.*

Démonstration. analogue à la démonstration pour la mesure de voisinage □

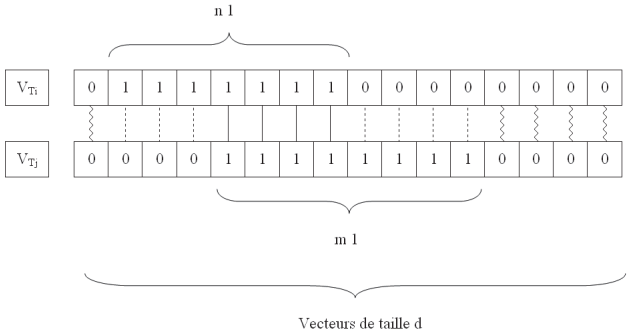
4.2.3 Croissance du nombre de mots

Théorème 4.6. *Pour des textes aléatoires le cosinus à la propriété suivante si $T_i \to T_j$ dans une constellation alors pour maximiser la cosinus entre les textes T_i et T_j il faut que le nombre de mots dans T_j soit plus grand que le nombre de mots dans T_i.*

Démonstration. Nous nous plaçons ici dans le cas de vecteurs booléens représentant les textes. Soit :

- V_{T_i} et V_{T_j} les vecteurs de deux textes T_i et T_j,
- d la taille de ces vecteurs,
- m le nombre de mots dans T_i i.e le nombre de composantes non nulles dans le vecteur V_{T_i},
- n le nombre de mots dans T_j i.e le nombre de composantes non nulles dans le vecteur V_{T_j}
- X le nombre de mots communs à T_i et T_j que l'on considérera comme aléatoire.

Nous cherchons la loi de X i.e du nombre de mots communs à T_i et T_j.
Si nous fixons le vecteur V_{T_i}, pour calculer la probabilité $P(X = k)$ i.e la probabilité qu'il y ait exactement k mots communs entre T_i et T_j, nous avons besoin du nombre de vecteurs V_{T_j} qui ont k mots communs avec V_{T_i} ainsi que le nombre de vecteurs V_{T_j} possibles :



nombre de vecteurs V_{T_j} qui a k mots communs avec V_{T_i}	C_m^k	on choisit pour le vecteur V_{T_j} k places parmi les m places non nulles du vecteur V_{T_i}
	C_{d-m}^{n-k}	comme on a choisi k places dans le vecteur V_{T_j}, on choisit dans le vecteur V_{T_j} n-k places parmi les d-m places nulles du vecteur V_{T_i}
le nombre de vecteurs V_{T_j} possibles	C_d^n	le nombre de vecteurs possibles V_{T_j} avec n composantes non nulles i.e le texte T_j contient n mots.

On a donc :

$$P(X = k) = \frac{C_m^k C_{d-m}^{n-k}}{C_d^n} \tag{III.11}$$

On reconnaît ici la loi hypergéométrique.
X suit une loi hypergéométrique de paramètre (d, m, n) et l'ensemble des possibles $X(\Omega)$ est l'ensemble des entiers entre $\max(0; n-(d-m))$ et $\min(m; n)$.
Par conséquent :

$$E(X) = \frac{nm}{d} \tag{III.12}$$

Le cosinus est une mesure de similarité c'est-à-dire que 2 textes seront d'autant plus proches qu'ils ont une mesure grande. Donc pour avoir des textes proches, il faut maximiser l'espérance. Pour V_{T_i} fixé i.e m fixé, pour maximiser E(X) il faut prendre

n le plus grand possible. Ce qui justifie que la similarité cosinus rapproche des textes de plus en plus grands. □

Nous avons construit une mesure de voisinage pour répondre au problème concret rencontré en veille stratégique en nous appuyant sur les propriétés suivantes :

- proposition d'un classement de textes en composantes : aux extrémités des composantes se trouvent les textes les plus longs et au centre (nucléus) les textes plus courts.
- la lisibilité des résultats et rapidité de lecture des textes,
- étiquetage automatique.

D'un point du vue pratique, nous avons développé un prototype, nommé Alhena, dans lequel nous avons implémenté notre mesure de voisinage. Les propriétés graphiques seront exploitées pour répondre à la problématique de l'équipe de veille stratégique. Le prototype est présenté au chapitre suivant.

Références bibliographiques

[BRN99] R. Baeza-Yates et B. Ribeiro-Neto : *Modern information retrieval*. 1999. 93

[BS08] M. Boughanem et J. Savoy : *Recherche d'information : état des lieux et perspectives*. Collection Recherche d'information et web, ISSN 1968-8008. Hermès science publ., Paris, 2008. 101

[CFLBC10] M.-L. Caron-Fasan, H. Lesca, A. Buitrago et A. Casagrande : Comment pérenniser un dispositif de veille anticipative à base de données numériques et textuelles : problématique et proposition. *In Actes colloque VSST'10*, 2010. 94

[CV04] Rudi Cilibrasi et Paul Vitanyi : Automatic meaning discovery using google. 2004. 96, 97

[CV06] R. Cilibrasi et P. Vitanyi : Automatic extraction of meaning from the web. pages 2309–2313, juillet 2006. 96

[CV07] R. Cilibrasi et P. Vitanyi : The google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, 19(3):370–383, 2007. 96

[LL11] H. Lesca et N. Lesca : *Les signaux faibles et la veille anticipative pour les décideurs : Méthodes et applications*. Hermes Science Publications, mai 2011. 94

[Mel01] S. Mellet : Lemmatisation et encodage grammatical : un luxe inutile ? *Lexicometrica*, (numéro spécial "Autour de la lemmatisation"), 2001. 93

[Tri07] A.-P. Trinh : Classification de texte et estimation probabiliste par machine à vecteur support. *Actes de l'atelier DEFT07, CAp07*, 2007. 93

Chapitre IV

Prototype Alhena

Dans ce chapitre, nous présentons le prototype Alhena qui est un outil d'aide à la navigation, basé sur notre mesure de voisinage. Nous montrons les différentes copies d'écrans que l'on rencontre lors de la navigation dans les graphes obtenus. Nous utilisons un corpus sur le thème « photovoltaïque ». Notre algorithme sera donné sous forme de pseudo-code et sa complexité algorithmique sera calculée.

Sommaire

Le prototype Alhena a été créé pour « opérationnaliser » la mesure de voisinage. Un des objectifs était de proposer un outil d'aide à la lecture des FULL texts en proposant des regroupements de FULL texts voisins. Nous avons décidé de nommer notre prototype Alhena en référence à une étoile de la constellation des Gémeaux car les graphes obtenus suggéraient un ciel étoilé de constellations.

L'animateur n'a pas besoin d'avoir un objectif précis (formuler son besoin sous forme de requête) lorsqu'il va utiliser Alhena. Il va le découvrir au fur et à mesure de son parcours dans le graphe proposé par Alhena. Nous présentons dans ce chapitre le prototype Alhena :

d'abord du côté utilisateur (section 1),

d'un point de vue informatique (section 2).

1 Alhena : côté utilisateur

Dans un premier temps, les textes sont insérés dans la base puis un algorithme de calcul de la mesure de voisinage et de traçage du graphe est lancé. Le graphe obtenu est appelé galaxie et les composantes de ce graphe sont nommées constellations locales (CL). Pincemin souligne que « *la valeur numérique de la mesure de similarité, que d'aucuns baptiseront "score de pertinence", sert essentiellement à sélectionner un nombre raisonnable de textes dans un corpus. Mais ce chiffre n'aide guère à percevoir quels aspects des textes ont motivé leur rapprochement. Des indications sont nécessaires, pour que les similarités calculées deviennent des rapprochements significatifs* »[Pin02]. Afin d'expliquer les rapprochements que nous proposons à l'aide de la mesure de voisinage, nous avons ajouté des informations dans les différents écrans fournis par Alhena.

1.1 Bases de textes et objectif de l'animateur

Nous avons constitué une base de 722 textes dont le sujet est le photovoltaïque. En 2011, 64 textes ont été collectés par des traqueurs et les 658 autres textes ont été

sélectionnés grâce à l'application Aproxima[1] [CFLBC10]. L'objectif de cette base était d'avoir un premier aperçu du domaine du photovoltaïque. Pour la séance de création collective de sens, l'animateur souhaite trouver des informations pouvant concerner directement ou indirectement l'entreprise française « Photowatt » qui fabrique des panneaux solaires.

1.2 Galaxie « photovoltaïque »

Dans un premier temps, les textes sont insérés dans la base de données d'Alhena. Ensuite le programme de calcul de la mesure de voisinage et de création de graphes est lancé. Dans le cas « photovoltaïque », on obtient la galaxie présentée à la figure IV.1. Cette galaxie est composée de 119 constellations locales. Comme expliqué au chapitre III, chaque texte a une flèche qui pointe sur le texte pour lequel la mesure de voisinage est minimale. Lorsqu'un texte T_i pointe sur un texte T_j, la flèche a une couleur qui dépend du calcul de la mesure de voisinage :

- la quantité de « rouge » est calculée en fonction du nombre de mots de L_{T_i} (liste des mots lemmatisés, inconnus et mal orthographiés de T_i) qui sont aussi dans L_{T_j},
- la quantité de « vert » est calculée en fonction du nombre de mots de L_{T_i} qui ont un synonyme dans L_{T_j},
- la quantité de « bleu » est calculée en fonction du nombre de mots de L_{T_i} qui ne sont pas présents dans L_{T_j}, qui n'ont pas de synonymes dans L_{T_j} et pour lesquels on a donc calculé la distance DCV.

Plus la flèche est rouge, plus les textes ont de mots en commun ; plus elle est verte plus les textes ont des synonymes. De plus le long de la flèche, nous avons indiqué un nombre qui exprime la valeur de la mesure de voisinage calculée entre les deux textes.

1. Approxima est un outil de collecte et de sélection des diverses informations numériques d'Internet ; il offre des fonctionnalités intéressantes de filtre et d'analyse d'informations brutes afin d'aider l'utilisateur à identifier les quelques informations susceptibles d'avoir un intérêt pour la phase d'exploitation des informations de veille et notamment des signaux faibles.

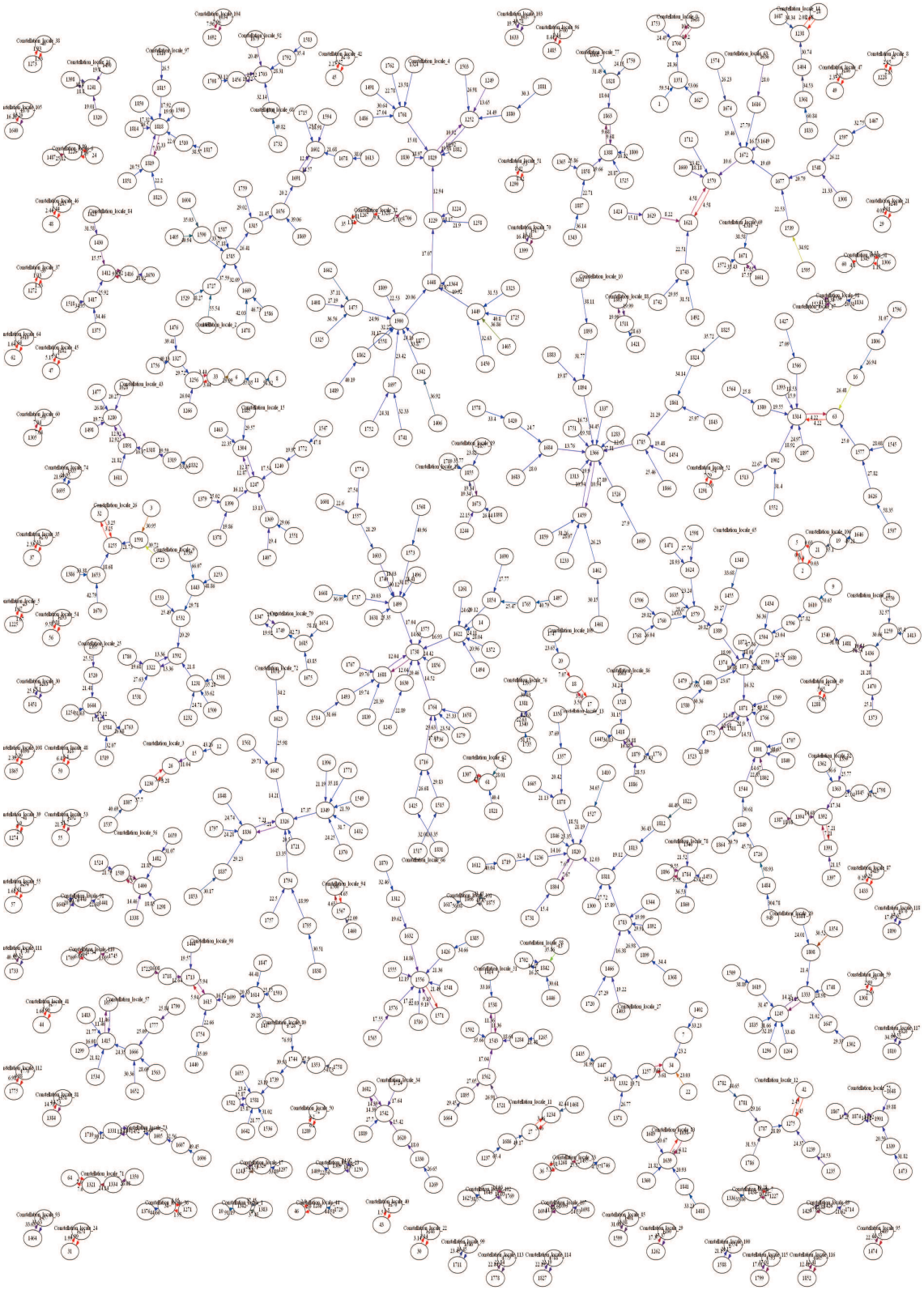


Figure IV.1 – Galaxie obtenue sur la base de textes « photovoltaïque »

1.3 Constellations locales

L'animateur va pouvoir regarder chacune des constellations locales en cliquant sur son nom, par exemple Constellation_Locale_84. Il obtient l'écran présenté à la figure IV.2. Cet écran contient la constellation locale (CL) ainsi que des nuages de mots :

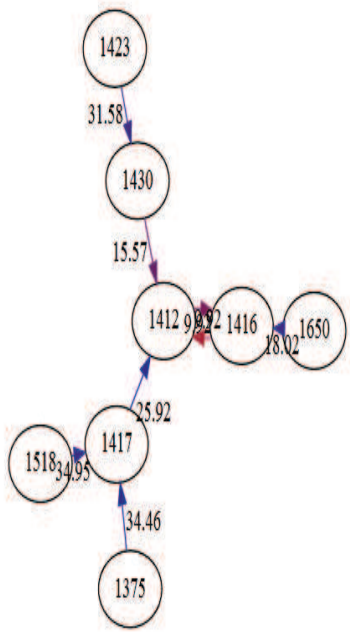
- le premier tableau contient le nuage de mots dans la constellation locale : plus les mots sont utilisés dans les textes de la CL, plus leur taille est importante (pour apparaître dans le nuage de mots, un mot doit être utilisé au moins deux fois)
- le deuxième tableau présente les nuages de mots de chacun des bras de la CL. Un bras part d'un texte vers lequel aucune flèche ne pointe et se termine dans le nucléus (ou cycle).

Dans notre exemple, la CL 84 de la galaxie « photovoltaïque » a dans son nuage de mots « Photowatt », terme qui intéresse l'animateur. De plus l'animateur remarque qu'un bras de la constellation semble aussi parler de manière importante de « Photowatt » : il est composé des textes 1518, 1417, 1412 et 1416.
Dans le chapitre III, nous avons montré que les bras des constellations locales étaient constitués de textes de plus en plus courts et qu'ainsi on va du général vers le particulier. L'animateur commence donc par explorer le nucléus. Il clique dans la CL sur le numéro du texte qui l'intéresse : donc par exemple le texte 1412.

1.4 Textes

L'animateur s'intéresse à la CL 84 et plus particulièrement à son nucléus composé des textes 1412 et 1416. En cliquant sur le numéro 1412, il obtient l'écran présenté à la figure IV.3. Cet écran donne à gauche le texte 1412 sur lequel on a cliqué et à droite le texte 1416 qui est le texte pour lequel la mesure de voisinage est minimale pour 1412. En haut du tableau se trouvent les nuages de mots de chacun des textes : comme pour le nuage de la CL, les mots les plus utilisés ont une taille plus importante et pour apparaître un mot doit être cité au moins deux fois dans le texte.
Lors du calcul de la mesure de voisinage, nous avons dû identifier :

- les mots communs de la liste de mots du texte 1412 et de celle de 1416,
- les mots de la liste de 1412 qui avaient un synonyme dans la liste de 1416,



énergie achat agence aider allemagne aller américain an année bal baser bauland bord bénéficier chine compétitif concurrence courbe coût créneau côtes-d'armor demander dur débat déjà écologie emploi enerplan euro figaro filière fonctionner fossile france industrie investir jean-louis mueth nucléaire photovoltaïque photowatt pionnier président solaire source temps territoire theréco thierry énergie

1412,1416,1417,1375,	énergie achat allemagne aller année bord compétitif courbe coût créneau dur déjà enerplan figaro filière fonctionner fossile france français frédéric gestion gouffre gouvernement grimper industrie industriel monicault mueth nucléaire ouest panneau pays photovoltaïque photowatt pionnier pouvoir prix production professionnel président question refonte rejoindre réseau solaire source tarif temps thierry énergie
1430,1412,1416,1423,	énergie agence aider américain an bal bauland bénéficier chine chinois concurrence créer demander difficulté écologie emploi filière france français gouvernement gw il illustrer indiquer industriel isère jean-louis marché mikkou molinois mondial ouest panneau philippe photovoltaïque photowatt presse prix problème production président secteur ser seul seulement société solaire souffrir électricité énergie
1518,1412,1416,1417,	'énergie allemagne attendre bord bour créneau daniel dur débat figaro français gouffre général industriel mesure moi|mois ouest panneau photovoltaïque photowatt pionnier pouvoir public représentant réunir sein société solaire solaire temps territoire élection énergie équipement état
1412,1416,1650,	'énergie baser côtes-d'armor entreprise euro france investir isère lézardrieux octobre ouest photowatt placer redressement société theréco

Figure IV.2 – Constellation locale 84 obtenue sur la base de textes « photovoltaïque »

Figure IV.3 – Ecran concernant le texte 1412

– les mots utilisés pour calculer la distance DCV minimale.

Nous avons choisi de faire apparaître ces informations de la façon suivante :

– les mots communs aux deux textes apparaissent surlignés en « saumon »,
– les mots synonymes écrits en bleu
– les mots pour lesquels la distance DCV a été calculée sont soulignés.

Dans notre exemple, on remarque que beaucoup de mots sont en commun et à la lecture, on note qu'il s'agit de la même information donnée par deux journaux différents. Pour plus de précision, l'animateur peut également regarder l'écran associé au texte 1416 (voir figure IV.4). L'animateur va ainsi pourvoir naviguer dans la galaxie « photovoltaïque ». Il a identifié la constellation locale 84 comme intéressante et l'exploration des autres constellations locales ne donnera rien d'intéressant.

1.5 Cas particuliers observables

Doublons Comme nous l'avons vu précédemment, les flèches dans la galaxie ont une couleur déterminée en fonction du nombre de mots communs, du nombre de synonymes et du nombre de mots pour lesquels on a calculé la distance DCV. Plus la flèche est rouge, plus les textes ont de mots communs. Deux textes d'un nucléus, la couleur des flèches qui tend fortement vers le « rouge » et une faible mesure de voisinage peuvent indiquer que les deux textes sont les mêmes à peu de différences près (différences de saisies). Par exemple dans la constellation « photovoltaïque »,

Figure IV.4 – Ecran concernant le texte 1416

les textes 1226 et 24 de la constellation locale 6 pointent l'un sur l'autre, les flèches sont rouges (voir figure IV.5) et la mesure de voisinage entre ces deux textes est de 1.89. Ils ont pour seule différence la manière dont a été saisi l'en-tête de l'article mais le corps de l'article est identique (voir figure IV.6).

Figure IV.5 – Ecran concernant la constellation locale 6

On peut aussi avoir le cas d'un même texte publié à deux dates différentes et/ou des sources différentes : il s'agit effectivement de doublons en ce qui concerne le contenu mais la différence de date peut avoir un intérêt pour la veille stratégique. On retrouve ce cas de figure pour les textes 1345 et 1306 (voir figure IV.7) qui contiennent le même texte mais publié à des dates différentes par des sources apparentes différentes.

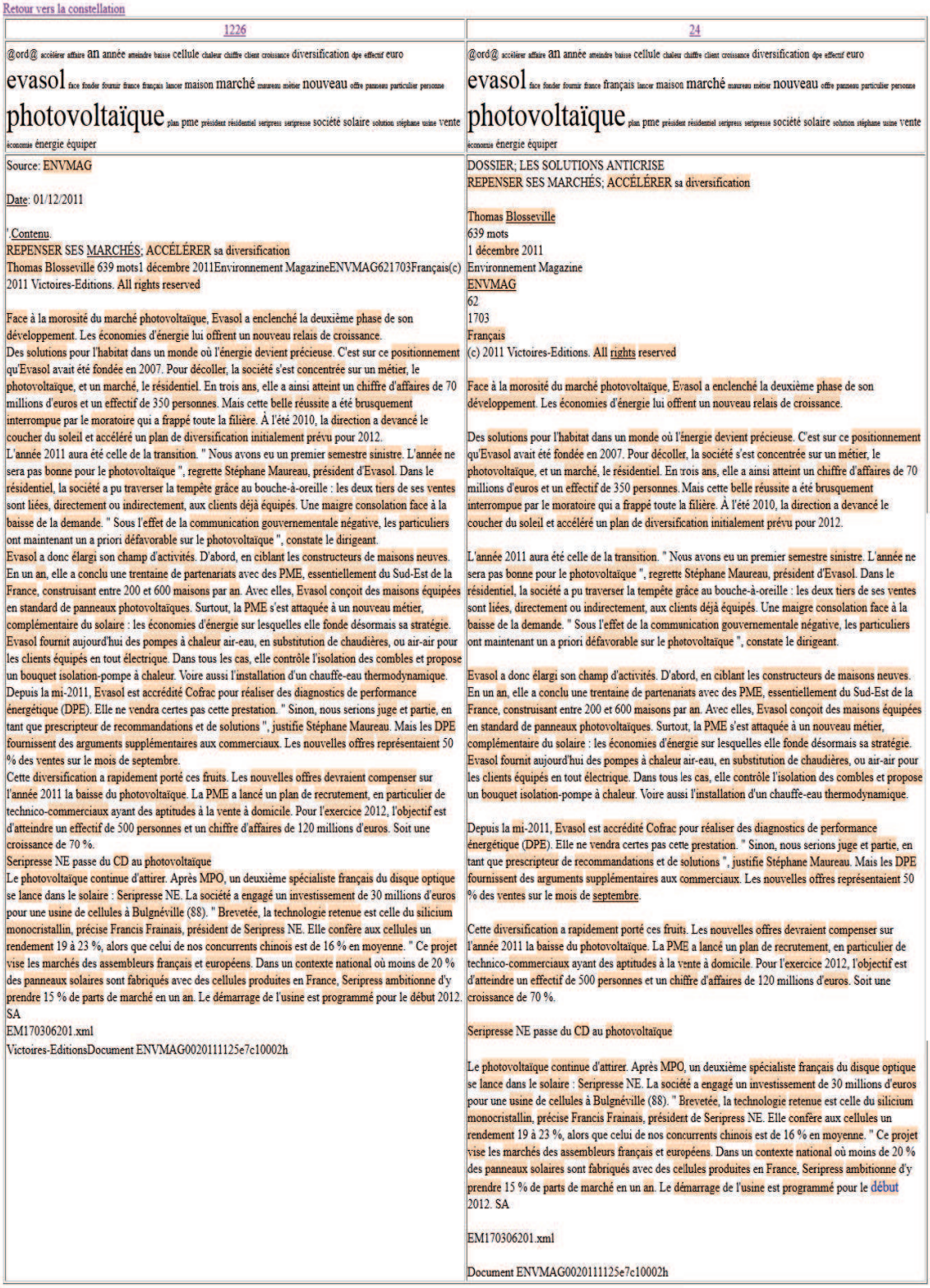


Figure IV.6 – Ecran concernant le texte 1126

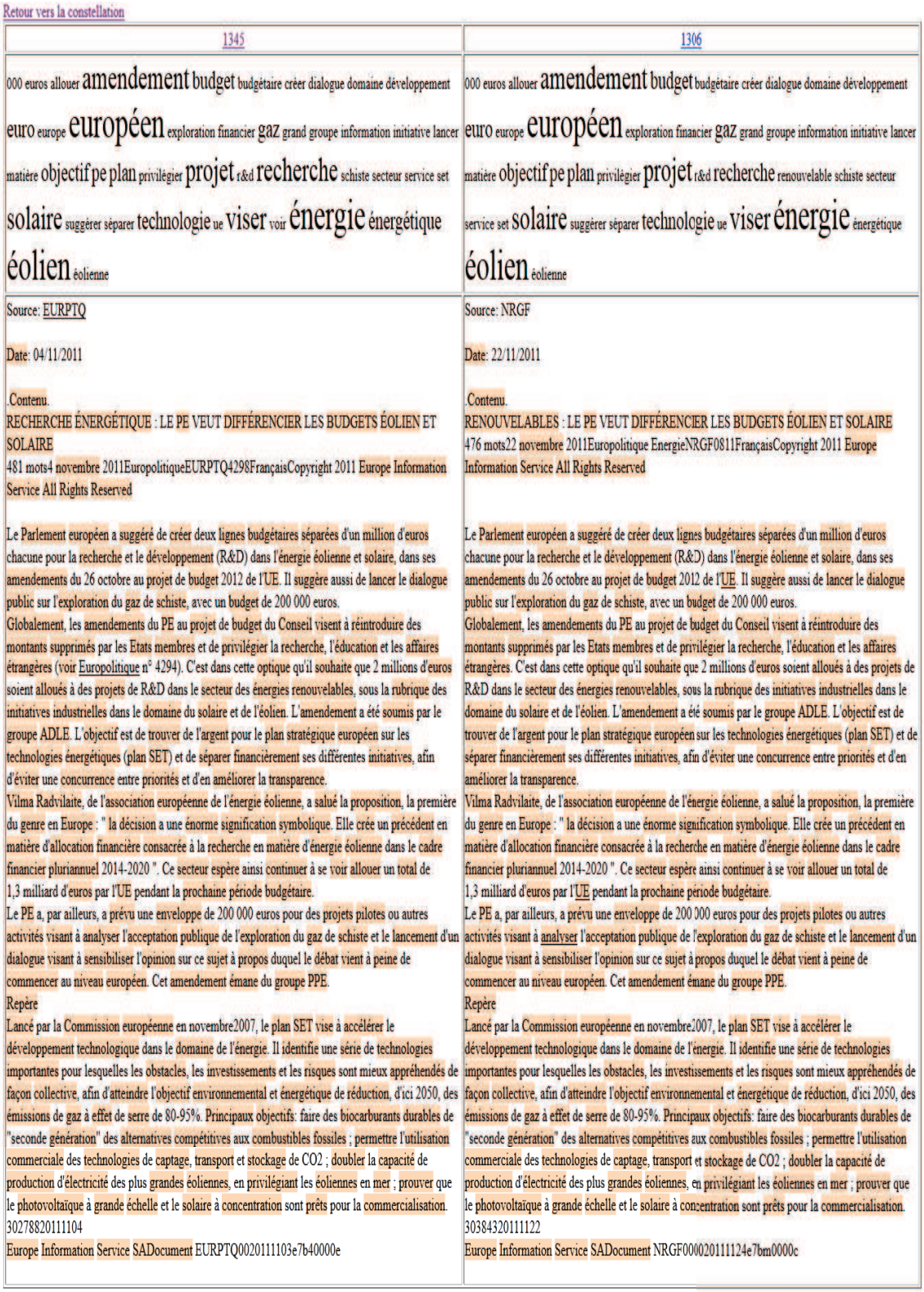


Figure IV.7 – Ecran concernant le texte 1345

Un texte inclus dans un autre Comme notre mesure de voisinage n'est pas normalisée par la taille des textes , elle permet d'associer deux textes dont l'un est contenu dans l'autre. Ce cas particulier est montré par la figure IV.8. On remarque que le texte 1384 est totalement inclus dans le texte 1358. Le texte 1358 a été publié deux jours plus tard et enrichi par rapport au texte 1384.

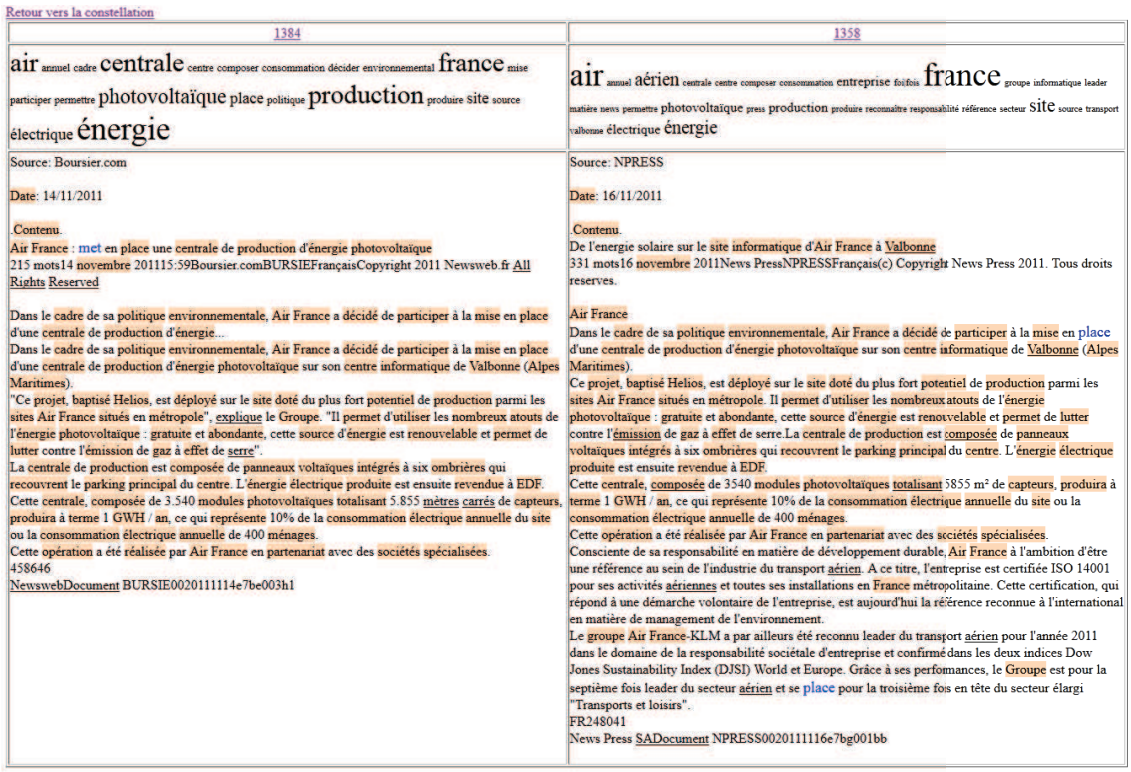


Figure IV.8 – Ecran concernant le texte 1384

Ces deux cas particuliers permettent d'aborder le problème de plagiat et de l'évolution dans le temps d'une informations. Pour conserver ces propriétés de « doublon » et d'« inclusion », nous avons choisi de ne pas normaliser notre mesure par la taille des textes.

2 Alhena côté informatique

2.1 Algorithme

Nous présentons maintenant l'algorithme en pseudo-code que nous avons implémenté dans Alhena pour calculer les mesures de voisinage entre les textes d'une base.

Données : les textes de la base avec leurs listes de mots lemmatisés, inconnus ou mal orthographiés
Résultat : une matrice contenant les mesures de voisinage entre les textes

pour chaque *texte* t_1 *de ma base* **faire**
  pour chaque *texte* t_2 *de ma base* **faire**
    $matriceMV[t_1, t_2] \leftarrow$ calculerMesureVoisinage(t_1,t_2)
    $matriceMV \leftarrow$ symétriser($matriceMV$)
  fin
fin

Algorithme 1: Programme principal

`Fonction calculerMesureVoisinage` (Liste de mots lemmatisés t_1, Liste de mots lemmatisés t_2)
début
  $mesure \leftarrow 0$
  pour chaque *mot lemmatisé* m_{t_1} *de* t_1 **faire**
    cas où m_{t_1} *appartient à* t_2
      mesure $\leftarrow$ mesure + 0 ;
    fin
    cas où m_{t_1} *a au moins un synonyme dans* t_2
      mesure $\leftarrow$ mesure + calculerMesureCasSynonyme(m_{t_1})
    fin
    autres cas mesure $\leftarrow$ $mesure$ + calculerDistanceDCV(m_{t_1},t_2)
  fin
  retourner *mesure*
fin

Algorithme 2: Fonction calculerMesureVoisinage

Données : Une table contient pour chaque mot lemmatisé, inconnu ou mal orthographié de la base la distance de Cilibrasi et Vitanyi la plus petite qu'il ait avec les autres mots lemmatisés de la base ; on note minDCV(mot) cette distance

`Fonction calculerMesureCasSynonyme` (Mot mot)
début
  $mesure \leftarrow \frac{minDCV(mot)}{2}$
  retourner *mesure*
fin

Algorithme 3: Fonction calculerMesureCasSynonyme

Notations : $nbTextes_{mot}$ =nombre de textes où apparait *mot*
$nbTextes_{mot_1,mot_2}$ =nombre de textes où apparait mot_1 et mot_2

`Fonction calculerDistanceDCV` (Mot mot, Texte t_2)
début
  $mesuremin \leftarrow +\infty$
  pour chaque *mot lemmatisé* m_{t_2} *de* t_2 **faire**
    $mesure \leftarrow \frac{\max(\log(nbTextes_{mot}),\log(nbTextes_{mot_{t_2}}))-\log(nbTextes_{mot,mot_{t_2}})}{\log(M)-\min(\log(nbTextes_{mot}),\log(nbTextes_{mot_{t_2}}))}$
    si $mesure < mesuremin$ **alors**
      $mesuremin \leftarrow mesure$
    fin
  fin
  retourner *mesuremin*
fin

Algorithme 4: Fonction calculerDistanceDCV

```
Fonction symétriser (Matrice matrice)
début
    pour chaque ligne de matrice faire
        pour chaque colonne de matrice faire
            matricesym[ligne,colonne] = (matrice[ligne,colonne]+matrice[colonne,ligne])/2
        fin
    fin
    retourner matricesym
fin
```

Algorithme 5: Fonction symétriser

2.2 Complexité

Nous calculons à l'aide des pseudo-codes la complexité de notre algorithme. Nous comptons le nombre de boucles imbriquées sur les données : r le nombre de mots et n le nombre de textes.

Algorithme	Complexité
Fonction « symétriser »	$\mathcal{O}(n^2)$
Fonction « calculerDistanceDCV »	$\mathcal{O}(r^2)$
Fonction « calculerMesureCasSynonyme »	$\mathcal{O}(r)$
Fonction « calculerMesureVoisinage »	$r * ($ complexité(« calculerMesureCasSynonyme »)+ complexité(« calculerDistanceDCV ») $= \mathcal{O}(r * (r + r^2)) = \mathcal{O}(r^3 + r^2)$
Programme principal	n^2* complexité(« calculerMesureVoisinage »)+ complexité(« symétriser ») $= \mathcal{O}(n^2(r^3 + r^2) + n^2) \Rightarrow \mathcal{O}(n^2 r^3)$

Tableau IV.1 – Calcul de la complexité

La complexité de notre algorithme est majorée par $\mathcal{O}(n^2 r^3)$. A titre de comparaison, la complexité de notre algorithme pour calculer le cosinus est majorée par $\mathcal{O}(n^2 r)$ car :

- le cosinus entre deux vecteurs : $\mathcal{O}(r)$,
- nous calculons le cosinus entre les n textes : $\mathcal{O}(n^2)$.

Nous avons montré dans ce chapitre les services offerts par notre prototype Alhena :

- une galaxie permettant une navigation dans les textes,
- une présentation de chaque constellation locale avec les nuages de mots associés à ces CL (nuage de mots de la CL, nuage de mots de chacun des bras de la CL),
- une page permettant de comprendre le rapprochement entre deux textes,
- un algorithme avec une complexité polynomiale en $\mathcal{O}(n^2r^3)$.

L'algorithme du calcul de la mesure de voisinage a une complexité polynomiale $\mathcal{O}(n^2r^3)$. Notre prototype Alhena doit maintenant être évalué au regard :

- de la veille stratégique : répondons-nous à la problématique de la recherche des informations voisines ? Apportons-nous une aide à l'animateur ?
- de l'informatique : la classification en constellation locale est-elle pertinente ? Peut-on mesurer cette pertinence ?
- d'autres domaines : peut-on utiliser Alhena pour d'autres domaines que la veille stratégique ? Quels sont les apports ?

Les réponses à ces questions font l'objet du chapitre suivant.

Références bibliographiques

[CFLBC10] M.-L. Caron-Fasan, H. Lesca, A. Buitrago et A. Casagrande : Comment pérenniser un dispositif de veille anticipative à base de données numériques et textuelles : problématique et proposition. *In Actes colloque VSST'10*, 2010. 124

[Pin02] B. Pincemin : Similarités texte texte. expérience d'une application de diffusion ciblée et propositions. *In Matematicas y Tratamiento de Corpus, Séminaire interlatin de linguistique appliquée*, page 35 52, 2002. 123

Chapitre V

Applications

Dans ce chapitre, nous montrons l'utilité du prototype Alhena en veille stratégique au travers de deux expériences :

- la première a été menée sur le thème de la valorisation du CO_2,
- la seconde sur le bioéthanol montrera un apport particulier d'Alhena dans le cadre d'un dispositif de veille : l'élargissement du champ de vision (« peripheral vision »).

Dans un deuxième temps, nous validons la pertinence de notre mesure en comparant les résultats obtenus avec Alhena et le classement « humain » d'une base d'articles de Reuters.

Nous montrons que la mesure de voisinage peut aussi être utilisée dans d'autres domaines :

- application en littérature : nous confronterons notre mesure aux Exercices de Style de Raymond Queneau.
- application en psychologie : nous présenterons le travail réalisé en collaboration avec des chercheurs en psychologie sur l'analyse de situations de mémoire.
- application en informatique : nous décrirons notre collaboration avec des chercheurs en informatique sur l'analyse de tchats.

Sommaire

Dans ce chapitre, nous présentons les différents cas où nous avons utilisé le prototype Alhena afin de valider nos travaux de recherche :

- une validation pour la veille stratégique : deux cas sont présentés le cas « CO_2 » et le cas « bioéthanol »
- une validation plus « informatique » : nous avons utilisé un corpus classé et nous avons comparé notre proposition de classement au classement proposé.

Notre prototype a également été utilisé en collaboration avec d'autres chercheurs dans deux domaines :

- en littérature sur les « Exercices de style »[Que47] de Raymond Queneau (collaboration avec Yolaine Cultiaux, chercheur associé à Pacte [1])
- en psychologie sur des exercices de classement de situations de mémoire (collaboration avec Fanny Vallet et Olivier Desrichard, chercheurs au LIP [2])
- en informatique sur des tchats enregistrés lors d'une expérience sur des jeux pédagogiques (travail en collaboration avec Jean-Charles Marty, chercheur au LIRIS [3] et Thibault Caron, chercheur au LIP6 [4]).

1 Applications à la veille stratégique

1.1 Cas : « Valorisation du CO_2 »

1.1.1 Présentation du cas

Dans le cadre d'un dispositif de veille stratégique, la direction de l'entreprise, appelée « Durability » (secteur de la chimie), a décidé d'explorer la possibilité d'exploiter le CO_2 en tant que ressource éventuelle (ou matière première), dans le but de diversifier ses activités dans une orientation stratégique d'avenir. Une séance de réflexion collective, réunissant divers directeurs (comité de direction), a été décidée. L'ordre du jour est libellé ainsi : « explorer la possibilité et la pertinence économique

1. Pacte est issue du regroupement de laboratoires dont les recherches portent sur la science politique, la géographie, l'aménagement et l'urbanisme, http ://www.pacte-grenoble.fr/
2. Le Laboratoire InterUniversitaire de Psychologie / Personnalité, Cognition, Changement Social (LIP/PC2S) est une unité de recherche consacrée à l'analyse de la cognition, du comportement et des interactions humaines dans leurs différents contextes, http ://www.lip.univ-savoie.fr/index.php
3. Laboratoire d'InfoRmatique en Image et Systèmes d'information, http ://liris.cnrs.fr
4. Laboratoire d'informatique de Paris 6, http ://www.lip6.fr/

de valoriser le CO_2 en tant que matière première éventuelle ». L'animateur est chargé de préparer les informations FULL texts susceptibles d'être utilisés au cours de la séance en fonction des interactions qui auront lieu entre les participants.
Dans une base de données comptant plusieurs centaines de textes, l'animateur a extrait 299 FULL texts pouvant correspondre à l'ordre du jour, mais il est hors de question que l'animateur les lise tous dans le peu de temps dont il dispose (surcharge d'information) : l'animateur doit donc extraire les plus pertinentes et leur nombre ne doit pas dépasser la quinzaine afin de les exploiter lors de la réunion.
Pour la réunion, l'animateur doit donc se mettre en condition :

- de présenter les FULL texts sélectionnés,
- de répondre rapidement aux demandes que pourraient formuler, « au fil de l'eau », les participants,
- d'accompagner le déroulement des interactions, sans briser le rythme de celles-ci, entre les participants en projetant les FULL texts éventuellement susceptibles d'aider à la réflexion collective,
- de répondre, au fur et à mesure et rapidement, à d'éventuelles questions du genre : « Cette information est-elle fiable ? Disposons-nous d'une information qui viendrait la compléter ? Disposons-nous d'une information qui viendrait contredire ou infirmer . . . ? ».

C'est dans le but d'apporter une aide efficiente, pour répondre à de telles conditions, qu'a été conçu et construit le prototype Alhena.
Au cours de la préparation de la future séance de travail, prévue pour le lendemain (pression du temps), la première tâche de l'animateur est de découvrir le contenu des 299 FULL texts. Il dispose de très peu de temps pour cela. Il utilise donc Alhena. Voici la suite des opérations qu'il a effectuées.

1.1.2 Résultats obtenus avec le prototype Alhena

Alhena affiche la représentation visuelle illustrée par la figure V.1 en forme de galaxie dans laquelle les 299 FULL text sont identifiables par leur numéro. L'animateur observe trois types de formes graphiques : des doubles flèches ; des petites constellations locales centrées sur leur nucléus (les doubles flèches) ; des bras des constellations locales, bras constitués des suites de FULL texts reliés entre eux par une simple flèche.

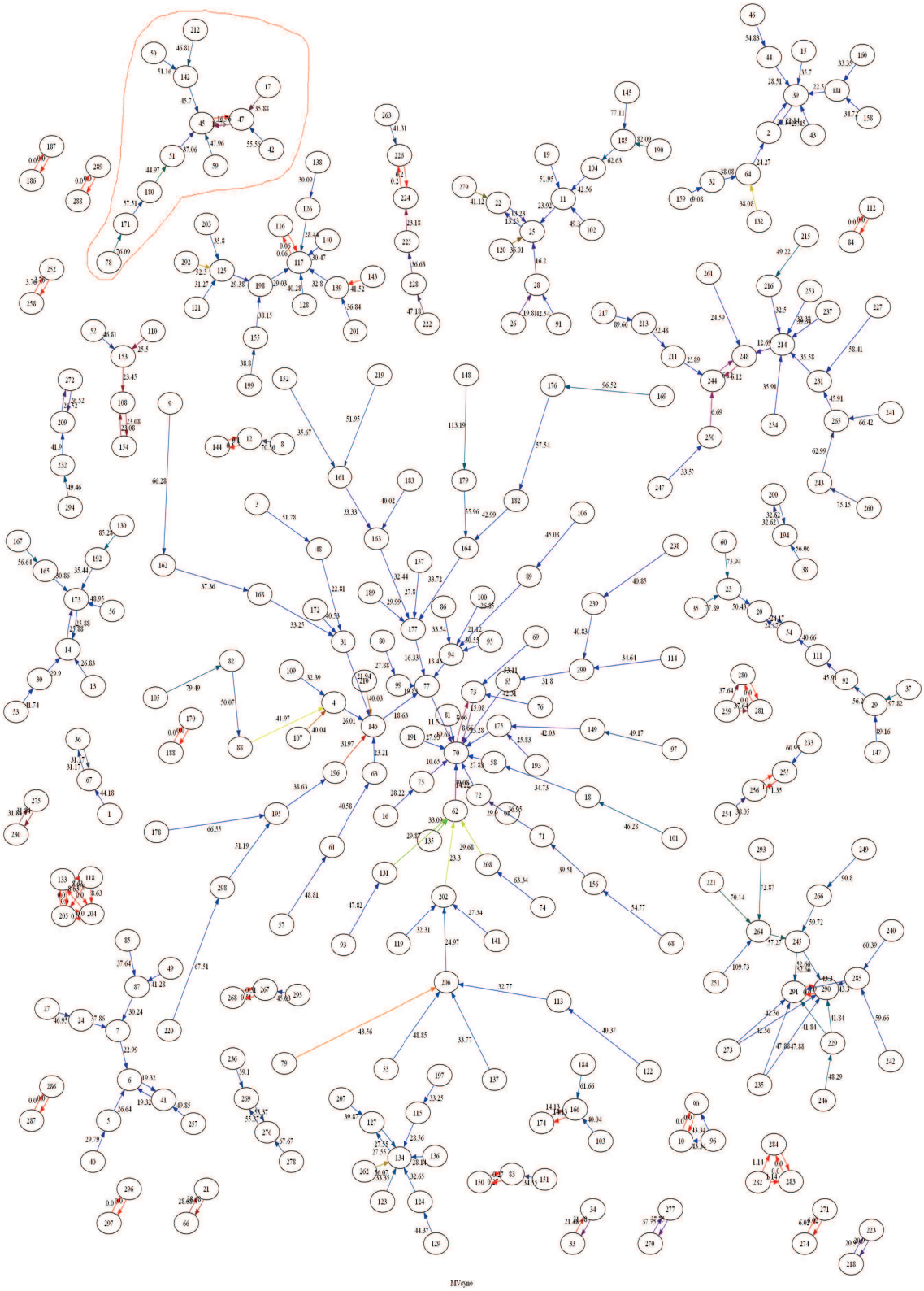


Figure V.1 – Galaxie obtenue avec Alhena sur la base CO_2

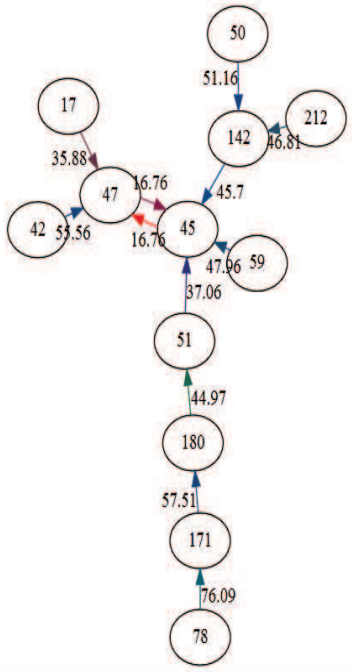
La démarche effectuée par l'animateur a été la suivante : pour commencer, l'animateur clique sur une constellation locale que nous noterons CL (par exemple celle entourée dans la figure V.1). Il obtient ainsi une page avec la représentation de la constellation locale, un tableau avec le nuage de mots de CL ainsi qu'un tableau contenant les nuages de mots pour chaque bras de la constellation. Le nuage de mots de CL permet à l'animateur d'avoir une idée générale sur le sujet abordé par CL. Les nuages de mots des branches donnent à l'animateur un aperçu de la manière dont le sujet de CL est abordé par chaque branche (voir figure V.2).

Analyse du nucléus Plusieurs conclusions émergent des constats que peut faire l'animateur :

- 45 et 47 ont un grand nombre de mots en commun mais ils ne sont pas des doublons car ils ne sont pas constitués du même nombre de mots (381 pour la FULL text 45 et 835 pour la FULL text 47) ; leurs sources d'émission ne sont pas les mêmes, pas plus que leurs dates d'émission (11 mars 2008 pour 47 et 18 mars pour 45). L'animateur peut conclure que 45 et 47 se fiabilisent mutuellement, ou tout au moins se confortent de façon significative. L'animateur sera ainsi en mesure de répondre à la question de la fiabilité si celle-ci lui est posée en cours de séance de travail collectif.
- deux noms d'acteurs apparaissent : Enerkem et GreenField, dont personne n'avait parlé jusqu'ici chez Durability. S'agirait-il de concurrents dont on n'avait pas pris conscience, concernant la piste stratégique explorée ?
- Durability explore la possibilité de valoriser le CO_2 en produisant de l'éthanol au moyen d'algues. Mais ici apparaît l'éthanol cellulosique : s'agirait-il d'une piste concurrente ... qui ferait appel à une technologie concurrente ? Faut-il s'en inquiéter ? L'attention est alertée sur des interrogations que les dirigeants n'avaient peut être pas encore envisagées (peripheral vision)
- La comparaison de 45 et 47 est largement facilitée puisque le logiciel surligne en couleur saumon tous les mots identiques dans 45, à gauche de l'écran et dans 47, à droite de l'écran.

En résumé,

- le logiciel attire l'attention sur un point d'entrée au sein des 299 FULL texts : le nucléus 45/47,



biocarburants biochimique biodiésel budget captage carburant cecr cellulosique charbon charge chef créer devoir dollar déchet enerkem estrie etat europe européen financement gaz greencentre greenfield greenline génération inc. industries innovation kingston leaf maroc norme papetier pays production produire produit recherche science stockage synthèse tourner tunis tunisie tunisien université usine étape éthanol

47,142,212,45,	accélérer agricole américain an avancer base biocarburants canada captage carburant cellulosique charbon chef co2 commercial concerner créer devoir diminuer département dépendance emploi enerkem exploitant gallon ges gouvernement greenfield groupe incitation interministériel litre montréal norme obama ontario production produire propre propre rapport renouvelable réduire société stockage stratégie technologie usine énergie éthanol
47,51,180,78,171,45,	arriver avancer biocarburants budget canada cecr cellulosique charge chef chimique chômage créer devoir dollar enerkem entreprise estrie etat europe européen exemple expertise exportation financement greencentre greenfield innovantes innovation jeune kingston maroc mondial pays politique production publix pôle rce recherche relance science technologique tourner tunis tunisie tunisien université usine éthanol étranger
47,45,17,	biocarburants biochimique carburant cellulosique chef chimique commercial conditionné déchet démarrage enerkem exploiter file gaz ges greenfield génération matière montréal ontario production produire produit québec renouvelable réduire situer société synthèse technologique transformer usine vert voiture westbury étape éthanol évoluer îlot
47,50,142,45,	américain an biocarburants bois canada cellulosique chef commercial concerner dollar déjà département enerkem emo fabriquer ges grand greenfield montréal ontario papetier papier pouvoir production propre rapport recherche réduire sembler société stora suède technologie usine énergie éthanol
47,59,45,	biocarburants canada capacité cellulosique chef commercial effectuer enerkem entreprise faibles ges greenfield installation montréal ontario organique partir permis production propre réacteur sherbrooke société technologie technologies téléphone urbain usine équipement éthanol
47,42,45,	biodiésel cellulosique commercial company eau ees energy enerkem gaz greenfield greenline inc. industries leaf montréal ontario production propre shaw société technologie usine énergie éthanol

Figure V.2 – Constellation locale 9 pour la base « CO_2 »

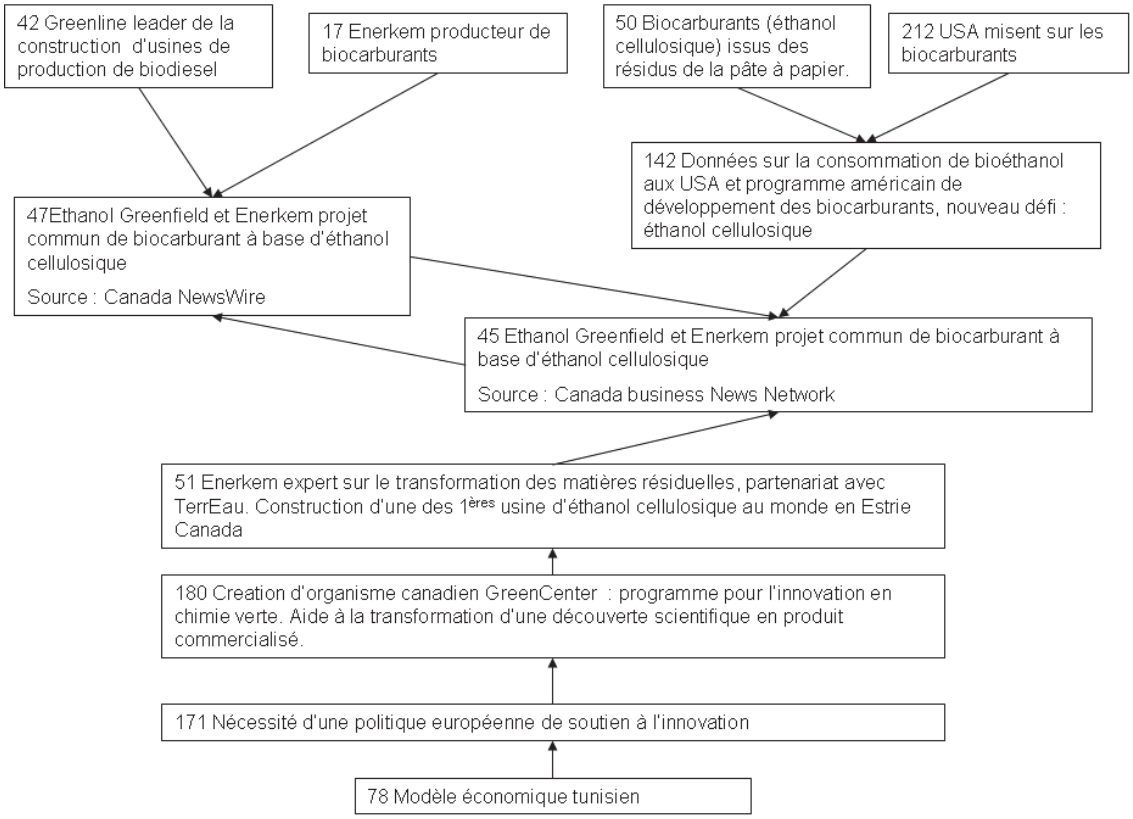


Figure V.3 – Extrait détaillé d'une constellation locale

- en pointant sur le nucléus, le logiciel suggère de nombreux éléments pour alimenter la réflexion stratégique du comité de direction de Durability,
- le temps nécessaire est très court comparé à ce qu'il serait si l'animateur avait dû tout faire manuellement, ce qui constitue un argument décisif aux yeux de la direction. Mais ce n'est pas le seul apport du prototype Alhena.

Petite couronne autour du nucléus L'animateur se demande si les informations 45 /47 peuvent être complétées, voire fiabilisées davantage. Il examine alors les FULL texts qui constituent une couronne rapprochée autour de 45/47 dans la constellation affichée sur l'écran, soit 51, 142, 17, 42 et 59. En cliquant tour à tour sur chacun de ces numéros, le texte complet du FULL text apparaît, de même que les mots les plus fréquents, dans la marge horizontale en haut de l'écran. L'animateur peut constater, en très peu de temps, que tous les textes de la couronne de 45 / 47 (sauf 42) apportent des éléments qui viennent compléter les renseignements

concernant les acteurs Enerkem et GreenField, d'une part, et l'éthanol cellulosique, d'autre part. Ces éléments pourront donc contribuer à rassurer les membres du comité de direction au sujet de la fiabilité des informations 45 /47. Ils pourront aussi susciter leur vigilance sur les concurrents potentiels Enerkem et Greenfield.

Bras de constellation L'animateur peut se demander pourquoi des FULL texts sont reliés entre eux pour constituer un bras de constellation locale. Par exemple 78, 171, 180, 51 constituent le bras le plus long de la constellation locale 45/47. Si l'animateur clique sur 78, Alhena affiche sur l'écran les textes complets des FULL texts 78 et de 171. Puis l'animateur peut cliquer sur 171, par exemple : apparaissent alors les textes de 171 et de 180, et ainsi de suite jusqu'à arriver au nucléus.

Pour les explications précédentes, nous avons choisi d'utiliser les informations 45/47, reliées par une double flèche et situées au centre d'une constellation locale (figure **??**). Nous avons pu montrer ainsi que l'existence d'une telle constellation locale est de nature à attirer l'attention sur une sous-thématique qui n'avait peut-être jamais été évoquée par les dirigeants de Durability jusque là, mais dont la découverte peut changer bien des éléments dans leur réflexion stratégique. Le « champ de vision » de la hiérarchie est augmenté là où il y avait un angle mort ! C'est une amplification de la vision périphérique collective que des auteurs appellent Peripheral Vision. Mais la galaxie contient d'autres constellations locales qui doivent être examinées de la même façon.

L'examen de la totalité des nucléus de la galaxie a été fait par l'animateur au cours de sa préparation. Le temps nécessaire a été d'environ deux heures : une heure de fonctionnement du logiciel (sans intervention humaine) puis une heure de « travail humain » pour l'analyse de la constellation par l'animateur. Le résultat est que seul un autre nucléus s'est également montré possiblement intéressant : le nucléus 70/73, il attire l'attention vers la piste « Industrie Alimentaire » pour la valorisation du CO_2. Par la suite, cette piste a surpris le comité de direction : il ne l'avait jamais évoquée jusqu'à là. Les questions stratégiques suivantes ont été soulevées : « Faut-il s'intéresser à cette piste ? Faudrait-il envisager des partenariats ? Quels pourraient être les débouchés dans la décennie à venir ? Etc. ».

La constellation globale évite de lire tous les FULL texts, ce qui serait impossible dans le temps disponible. Les nucléus permettent d'aller à l'essentiel du groupe de FULL texts constituant une constellation locale. L'animateur peut découvrir très vite si l'ensemble des FULL text constituant une constellation locale est pertinent ou non au regard de l'ordre du jour.
Les informations voisines d'un FULL text fournissent très rapidement des éléments pour répondre aux questions telles que : cette information est-elle fiable ? Pouvons-nous la compléter un peu ? Cette information est-elle en contradiction avec une autre information ? Avons-nous découvert une sous-problématique à laquelle nous n'avions pas pensé jusque là (peripheral vision) ? Toutes ces questions ont reçu des réponses très rapidement et compatibles avec le bon déroulement des interactions entre les participants.

1.2 Ouvrir la réflexion stratégique « Peripheral Vision » : cas du « bioéthanol »

Dans ce cas, la problématique indiquée par la hiérarchie et devant être examinée par la séance de Création Collective de Sens (CCS) est la suivante : « Comment valoriser le CO_2 comme matière première ? Devrions-nous développer une nouvelle activité dans le domaine des biocarburants et plus précisément le bioéthanol ? ».

1.2.1 Expérimentation sans Alhena

A L'animateur doit préparer la séance de CCS qui doit avoir lieu le lendemain. Il a recueilli les données qu'il a pu trouver auprès de ses sources d'information. Celles-là sont maintenant stockées dans sa base de données. De chacun des FULL texts, il a extrait quelques mots significatifs qui constituent ce que nous appellerons des brèves. Celles-ci pourront être projetées sur l'écran situé dans la salle de travail. Le texte complet relatif à chacune des brèves sera disponible dans la salle de travail. Parmi elles figure, par exemple, le FULL text 93 (nous écrirons FULL93 par la suite).

B Au cours de la séance de CCS (donc le lendemain) les participants, membres du Comité d'exploitation des informations (CODEXI), examinent les brèves proposées par l'animateur, chacune à leur tour. Les participants échangent leurs points de vue au cours d'un débat délibératif. S'agissant de la brève 93, plusieurs participants éprouvent le besoin d'en savoir un peu plus. Ils demandent à l'animateur de rechercher des données pouvant éclairer la brève 93.

C L'animateur interroge sa base de données au moyen de mots-clés. Quels mots clés choisir dans le texte FULL 93 ? Les participants indiquent qu'il conviendrait de savoir ce que peut leur apporter le « Rapport 2009 », d'une part et les données concernant « éthanol cellulosique », d'autre part. L'animateur effectue la recherche et obtient ainsi :

- le « Rapport 2009 », qui compte 87 pages, et
- une douzaine de FULL texts apportant des précisions sur « éthanol cellulosique » : la description de celui-ci, comment il est obtenu, etc.

D Un participant jette un coup d'œil sur le rapport en présence des autres membres du CODEXI. Mais tous déclarent qu'il est hors de question d'exploiter ce volumineux rapport au cours de la présente séance prévue pour durer deux heures ! Les autres FULL texts sont rapidement parcourues du regard : elles ne semblent pas apporter grand-chose pour les réflexions en cours.

E Les participants continuent d'échanger des réflexions sur la base des brèves apportées en début de séance. In fine, ils concluent qu'ils n'en savent pas assez sur la thématique à l'ordre du jour : une autre séance sera prévue, lorsque l'animateur aura trouvé des données plus pertinentes. Quelques indications lui sont données pour qu'il puisse orienter ses recherches.

1.2.2 Expérimentation avec Alhena

Les étapes A et B sont identiques à celles décrites ci-dessus.

C' L'animateur va utiliser le logiciel Alhena sur la totalité de sa base de données FULL text. Il n'indique aucun mot clé. Suite à la recherche sur ordinateur déclen-

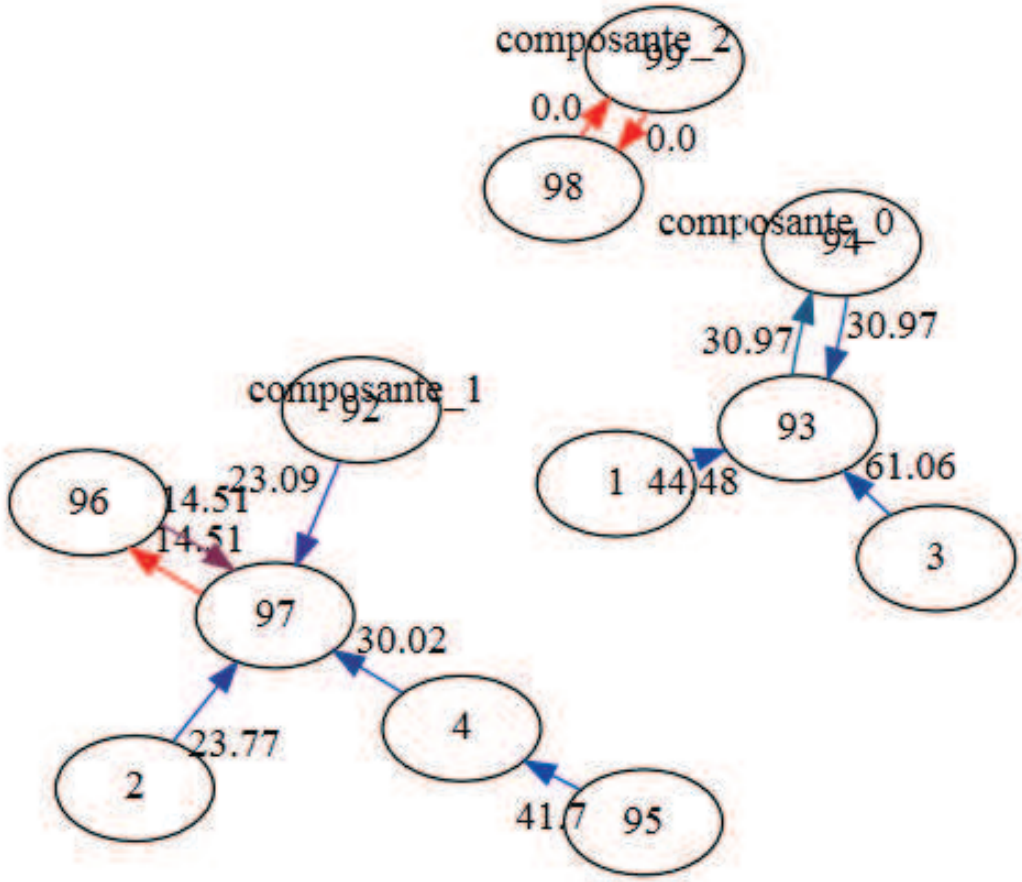


Figure V.4 – Galaxie liée à la base Bioéthanol

chée par l'animateur, Alhena affiche sur l'écran une galaxie (figure V.4). L'animateur propose : « Portons notre attention sur la constellation dont le centre est FULL93 ». Il apparaît que FULL93 est entouré de trois FULL numérotés 1, 3 et 94. Nous dirons qu'il s'agit de trois FULL « voisines » de FULL93.
Question : ces trois « informations voisines » viennent-elles enrichir FULL93 ? Viennent-elles élargir le champ de vision de l'animateur (et des participants CCS si la recherche est faite en leur présence) ? Que constatent alors les participants, à qui les trois FULL sont distribués ?

La lecture de FULL1 leur apprend : l'existence d'une société nommée SAPPHIRE. Celle-ci est bien avancée vers la production de biocarburants à base d'algues. Les participants ignoraient (peut-être) l'existence de celle-ci. Bill Gates est sponsor

de SAPPHIRE : c'est signe de l'importance potentielle de cette société. SAPPHIRE connaît une croissance d'une rapidité étonnante (des chiffres sont mentionnés). Ces éléments d'information n'étaient pas contenus dans FULL93. Les participants découvrent des informations « à la périphérie de FULL93 » qui viennent l'enrichir. En d'autres termes, la « vision périphérique » (peripheral vision) de chacun des participants est augmentée.
La lecture de FULL3 leur apprend que : L'existence de la mise au point d'une nouvelle méthode de conversion directe d'algues humides en biodiesel ; mise au point réalisée par un laboratoire de l'université du Michigan. Il est possible que le biodiesel n'intéresse pas le CODEXI Mais c'est lui qui en décidera. Le « laboratoire Michigan » apparaît comme un acteur pouvant susciter l'attention : partenaire éventuel ? concurrent ? etc. Il sera possible de joindre ce laboratoire afin d'en savoir plus, si cela est nécessaire. Le champ de vision autour de FULL93 est de nouveau élargi.

La lecture de FULL94 : Apprend qu'il existe une agence nommée RAIB dont les recherches sont orientées vers la production d'ammoniac à l'état sec. Les participants décideront si cet ammoniac concerne leur sujet d'intérêt. Peut-être l'agence RAIB est-elle actuellement inconnue du CODEXI ? En tout cas, l'adresse mail de son site officiel est maintenant connue. Ce site est en japonais ... ce qui donne à penser que des recherches sont effectuées au Japon, probablement par divers laboratoires : lesquelles ?

Si l'utilisation de Alhena est faite par l'animateur en amont de la séance de CCS, celui-ci sera mieux armé pour préparer la séance, disposant d'informations (FULL texts) à la fois peu nombreuses et pertinentes pour répondre aux demandes à lui adressées. L'animateur voit son rôle valorisé au regard du CODEXI et de la hiérarchie.

Cet exemple montre l'utilité de Alhena pour élargir le champ de vision périphérique autour d'une information peu signifiante si elle était isolée, sans « voisines ».

2 Résultats sur une base classée : Reuters

2.1 Présentation de la base

Afin d'évaluer la pertinence de notre mesure de voisinage, nous avons utilisé la base de textes Reuters-21578 [5][Yan99][Seb02][MB04]. Plus précisément, nous avons utilisé les 10 top-classes de ce corpus. Le tableau V.1 présente les classes de Reuters (certains textes appartiennent à plusieurs classes).

Classes	Nombre de textes
acq	719
corn	56
crude	189
earn	1087
grain	149
interest	131
money-fx	179
ship	89
trade	117
wheat	71
Nombre de textes différents	2545

Tableau V.1 – 10 top-classes de Reuters-21578

2.2 Résultats

Alhena a calculé les mesures de voisinage entre les 2545 textes et a construit le graphe des minimums présenté à la figure V.5. Cette galaxie contient 368 constellations locales (CL).
Afin d'évaluer notre mesure, nous avons conservé de la matrice des mesures de voisinage la première mesure minimale (min1) et la deuxième mesure minimale (min2). Ensuite nous avons comptabilisé sur la base totale le nombre de fois où le min1 met en correspondance deux textes de la même classe (tableau V.2) c'est-à-dire le nombre de fois où notre mesure est en adéquation avec le classement de Reuters par les humains. Nous avons fait de même avec le min2 et avec à la fois les min1 et min2

5. Reuters-21578 Distribution 1.0 est disponible sur le site http ://disi.unitn.it/moschitti/corpora.htm

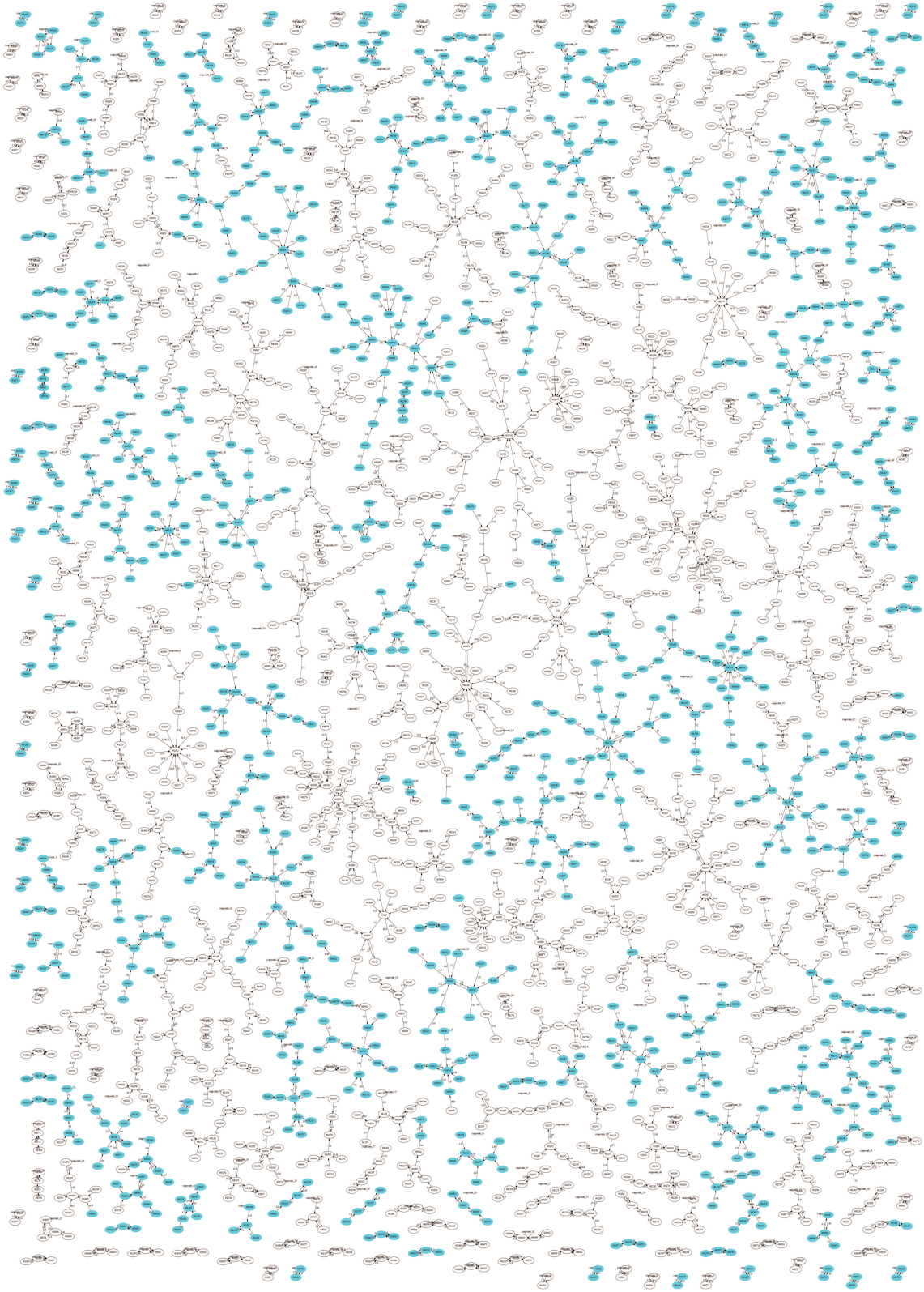


Figure V.5 – Galaxie obtenue avec Alhena sur la base Reuters avec en bleu les textes de la classe "earn"

Condition	Nombre de textes vérifiant la condition	%	Nombre de textes ne vérifiant pas la condition	%
tous les min1 sont dans la même classe	2329	91,51%	216	8,49%
tous les min2 sont dans la même classe	2249	88,37%	296	11,63%
tous les min1 et min2 sont dans la même classe	2152	84,56%	393	15,44%

Tableau V.2 – Statistiques

pour tester la stabilité des classes. Ensuite nous avons réalisé les mêmes calculs par classe (figure V.6).

Comme on l'a vu au chapitre II, en recherche d'information, la précision et le rappel sont les outils classiques d'évaluation des outils :

- le rappel est défini par le nombre de documents pertinents retrouvés par rapport au nombre de documents pertinents de la base,
- la précision est le nombre de documents pertinents retrouvés par rapport au nombre total de documents donnés par l'outil.

Dans notre cas :

$$Rappel_{classe_i} = \frac{Nombre\ de\ textes \in classe_i\ dont\ le\ min1\ \in classe_i}{Nombre\ de\ textes \in classe_i} \tag{V.1}$$

$$Précision_{classe_i} = \frac{Nombre\ de\ textes \in classe_i\ dont\ le\ min1\ \in classe_i}{Nombre\ de\ textes\ attribués\ à\ la\ classe_i} \tag{V.2}$$

Le tableau V.3 présente les calculs de la précision et du rappel.

Dans près de 92% des cas, notre mesure minimale permet d'associer un texte d'une classe à un texte de la même classe. Lorsque l'on regarde les résultats par classe (% de « bien classés », rappel et précision), les meilleurs scores sont obtenus sur les

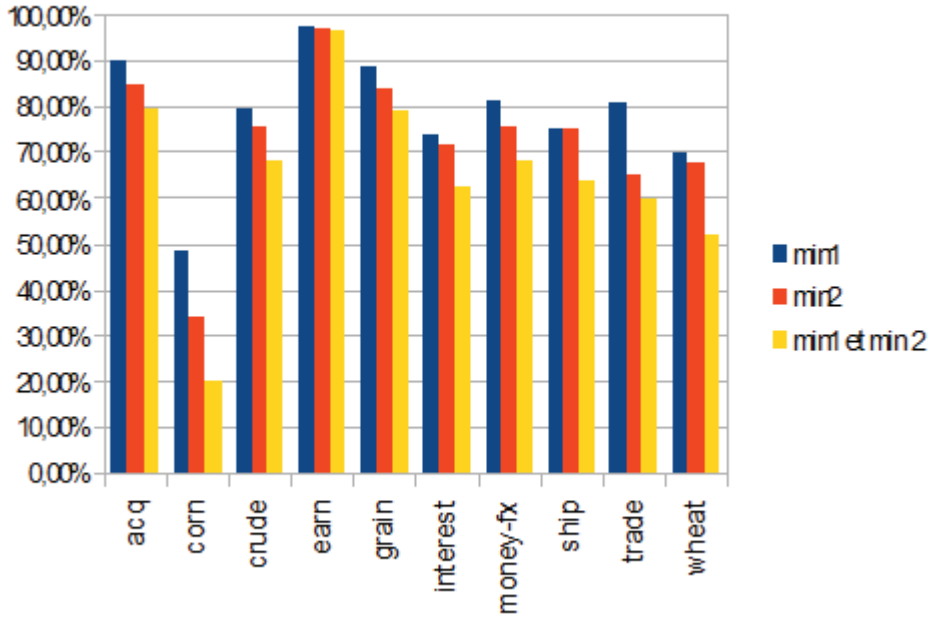


Figure V.6 – Statistiques par classe sur les résultats obtenus avec la mesure de voisinage

Classe	Rappel	Précision
acq	0,8998609179	0,8998609179
corn	0,4821428571	0,4576271186
crude	0,7989417989	0,7823834197
earn	0,9797608096	0,9761686526
grain	0,8859060403	0,862745098
interest	0,7404580153	0,723880597
money-fx	0,8156424581	0,8066298343
ship	0,7528089888	0,7282608696
trade	0,811965812	0,7983193277
wheat	0,7042253521	0,6849315068
total	0,787171305	0,7720807342

Tableau V.3 – Rappel et Précision

classes les plus importantes en taille. Notre plus mauvais score concerne la classe « corn ». En effet, cette classe contient peu de textes et l'ensemble de ces textes a été également classé dans d'autres classes par les humains : ce qui explique les mauvais résultats.

Afin de valider la pertinence de notre mesure de voisinage, nous avons procédé de la même manière avec le cosinus (figure V.7). Nous avons également testé si une différence significative existait pour chaque classe entre la proportion de bien classés (min1) obtenue avec la mesure de voisinage et celle obtenue avec le cosinus ; pour cela nous avons réalisé des tests d'égalité des proportions [6] avec un risque α de 5%. Nous obtenons les résultats présentés au tableau V.4. Au vu des résultats, nous pouvons conclure que notre mesure de voisinage propose un classement obtenant de bons scores de rappels et précisions. En outre, notre mesure classe aussi bien que le cosinus. La différence est que notre mesure propose une « analyse » de textes (du général au particulier) alors que le cosinus propose une « synthèse » (du particulier vers le général).

6. le descriptif du test est donné en annexes

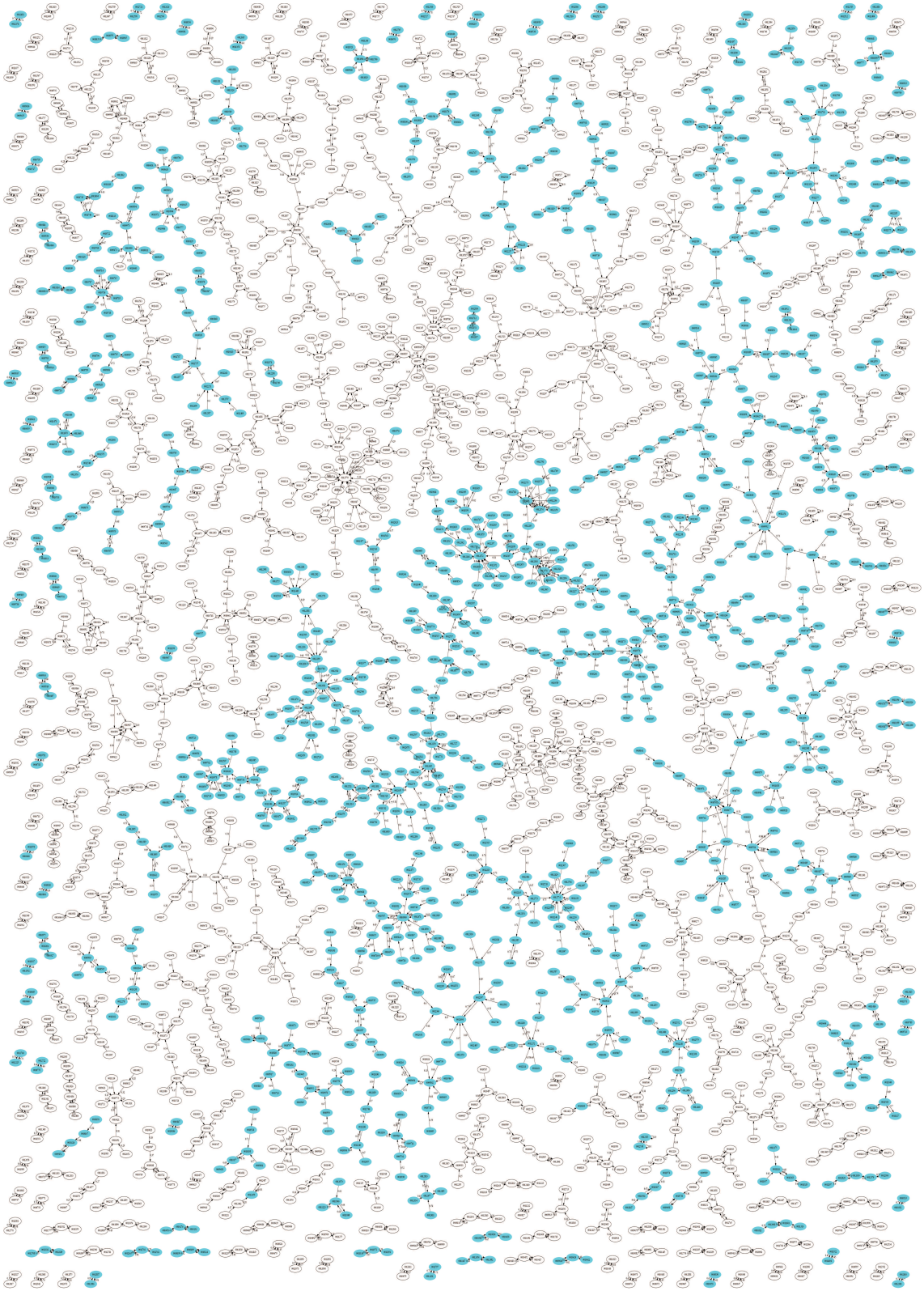


Figure V.7 – Galaxie obtenue avec le cosinus sur la base Reuters avec en bleu les textes de la classe "earn"

Classe	Nombre de textes dans la classe	Nombre de textes bien classés avec la mesure de voisinage (min1)	Nombre de textes bien classés avec le cosinus (min1)	Résultat du test
acq	719	647	656	pas de différence significative
corn	56	25	30	pas de différence significative
crude	189	147	161	pas de différence significative
earn	1087	1061	1073	pas de différence significative
grain	149	128	134	pas de différence significative
interest	131	94	102	pas de différence significative
money-fx	179	144	146	pas de différence significative
ship	89	64	74	pas de différence significative
trade	117	93	100	pas de différence significative
wheat	71	48	53	pas de différence significative

Tableau V.4 – Comparaison des résultats obtenus avec la mesure de voisinage et le cosinus

3 Exercices de style de Raymond Queneau

Il est difficile d'analyser automatiquement des textes littéraires utilisant des figures de style comme des métaphores ou des métonymies. Nous avons décidé de confronter notre mesure aux « Exercices de style »[Que47] de Raymond Queneau qui offrent une grand variété stylistique.

3.1 Présentation

Queneau a écrit en 1947 « Exercices de style » : l'auteur raconte de 99 façons différentes la même histoire :
« Un voyageur attend le bus, il remarque un jeune homme au long cou qui porte un chapeau bizarre, entouré d'un galon tressé. Le jeune homme se dispute avec un passager qui lui reproche de lui marcher sur les pieds chaque fois que quelqu'un monte ou descend. Puis il va s'asseoir sur un siège inoccupé. Un quart d'heure plus tard le voyageur revoit le jeune homme devant la gare Saint-Lazare. Il discute avec un ami à propos d'un bouton de pardessus. »
Queneau s'est imposé à chaque fois une contrainte stylistique, ce qui donne par exemple :
« Métaphoriquement.
Au centre du jour, jeté dans le tas des sardines voyageuses d'un coléoptère à grosse carapace blanche, un poulet au grand cou déplumé harangua soudain l'une, paisible, d'entre elles et son langage se déploya dans les airs, humide d'une protestation. Puis attiré par un vide, l'oisillon s'y précipita.
Dans un morne désert urbain, je le revis le jour même se faisant moucher l'arrogance pour un quelconque bouton. »
ou
« Précisions.
Dans un autobus de la ligne S, long de 10 mètres, large de 2,1, haut de 3,5, à 3 km. 600 de son point de départ, alors qu'il était chargé de 48 personnes, à 12 h. 17, un individu de sexe masculin, âgé de 27 ans 3 mois 8 jours, taille de 1 m 72 et pesant 65 kg et portant sur la tête un chapeau haut de 17 centimètres dont la calotte était entourée d'un ruban long de 35 centimètres, interpelle un homme âgé de 48 ans 4 mois 3 jours et de taille 1 m 68 et pesant 77 kg., au moyen de 14 mots dont l'énon-

ciation dura 5 secondes et qui faisaient allusion à des déplacements involontaires de 15 à 20 millimètres. Il va ensuite s'asseoir à quelque 2 m. 10 de là.
118 minutes plus tard il se trouvait à 10 mètres de la gare Saint-Lazare, entrée banlieue, et se promenait de long en large sur un trajet de 30 mètres avec un camarade âgé de 28 ans,taille 1 m. 70 et pesant 71 kg. qui lui conseilla en 15 mots de déplacer de 5 centimètres, dans la direction du zénith, un bouton de 3 centimètres de diamètre. »
ou encore
« Interjections.
Psst ! heu ! ah ! oh ! hum ! ah ! ouf ! eh ! tiens ! oh ! peuh ! pouah ! ouïe ! ou ! aïe ! eh ! hein ! heu ! pfuitt ! Tiens ! eh ! peuh ! oh ! heu ! bon ! »
Nous avons constitué une base contenant les 99 exercices de style et le texte de présentation.

3.2 Résultats

Nous avons mené une expérience en deux temps avec les textes de Queneau :

- étape 1 : évaluer les différences entre un classement « humain » et un classement par Alhena. Nous avons demandé à Yolaine Cultiaux, spécialiste de sciences politiques et de littérature, de lire et de classer les textes de Queneau en fonction du vocabulaire et nous avons comparé ce classement avec la galaxie obtenue avec Alhena.
- étape 2 : Yolaine Cultiaux a parcouru la galaxie et a essayé de donner une « définition » à chaque nucléus, chaque bras et chaque composante locale.

La figure V.8 présente la galaxie obtenue sur les textes en français de Queneau.

Étape 1

Nous avons demandé à Yolaine Cultiaux de classer les exercices de style en fonction du vocabulaire. La classification qui nous a été proposée est présentée dans le tableau V.5. La figure V.9 présente la galaxie obtenue avec Alhena avec en couleur la classification proposée par Yolaine Cultiaux. A la suite d'un échange sur les classifications obtenues, nous avons noté que :

- le temps pour lire et classer les textes avait été très important (5 à 6 heures

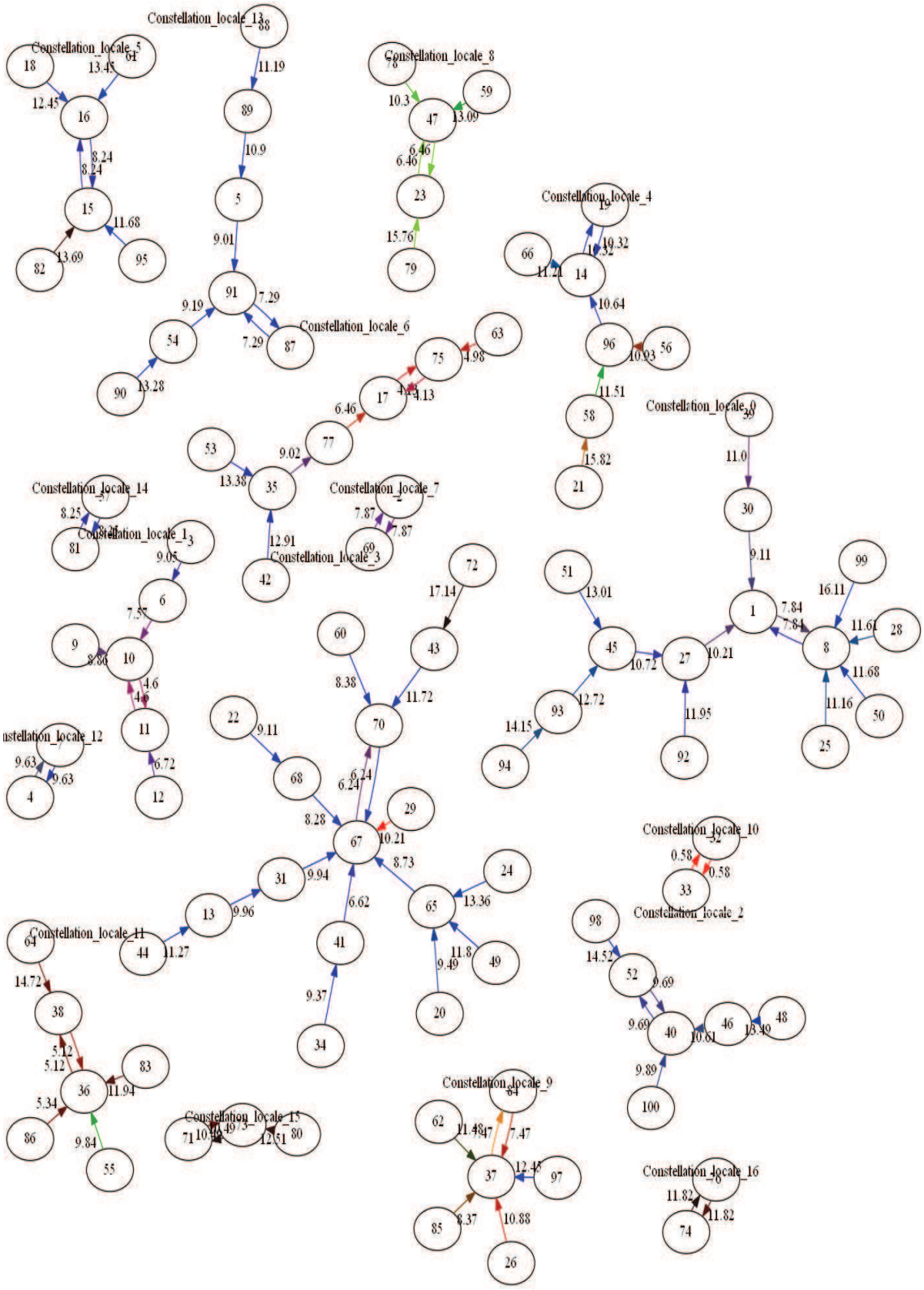


Figure V.8 – Galaxie obtenue avec Alhena sur la base des textes de Queneau

Nom de la classe	Textes
Récits précis	1, 2, 3, 4, 6, 7, 9, 10, 12, 14, 15, 16, 17, 20, 24, 28, 30, 32, 33, 34, 35, 39, 40, 41, 42, 44, 45, 48, 52, 53, 60, 61, 64, 66, 75, 81, 87, 88, 89, 90, 91, 92, 93, 95, 99, 100
Textes imprécis	8, 13, 19, 22, 25, 27, 28, 29, 35, 39, 41, 46, 49, 51, 52, 55, 56, 57, 58, 59, 64, 65, 66, 68, 69, 70, 71, 77, 78, 88, 92, 94, 95, 96, 99, 100
Subjectivité	2, 3, 7, 8, 10, 15, 17, 20, 24, 29, 39, 40, 46, 53, 66, 67, 68, 87, 88, 89, 90, 91, 92, 93, 94, 97, 99
Partial négatif	3, 7, 9, 34, 40, 42, 44, 53, 60, 66, 67, 70, 87, 88, 89, 90, 91, 92, 93, 95, 97, 99, 100
Neutre	75
Poésie	5, 23, 31, 42, 47, 50, 54, 61, 64, 67, 68, 83, 87, 91, 92, 95, 99
Sensoriel	11, 55, 56, 57, 58, 59, 90
Familier	12, 16, 26, 34, 39, 43, 45, 50, 54, 57, 58, 60, 61, 70, 72, 73, 74, 76, 81, 83, 87, 90, 91, 93, 95, 97, 98
Langage soutenu	31, 32, 42, 48, 49, 99
Jeux avec les mots	12, 18, 21, 36, 37, 38, 47, 62, 63, 65, 69, 83, 86, 96
Méta	25, 27, 50, 51, 69, 95
Incompréhensible	71, 72, 73, 74, 76, 79, 80, 82, 83, 84, 85, 97, 98

Tableau V.5 – Classification « humaine » des exercices de style

réparties sur plusieurs jours),
- le classement de Yolaine Cultiaux et les critères de classement avaient évolué au fur et à mesure de la lecture,
- le classement de Yolaine Cultiaux était aussi lié à la sémantique (mot subjectif, jeux de mots, ...) contrairement au classement Alhena qui s'appuie uniquement sur la graphie des mots et non sur leur sens.
- un texte peut appartenir à plusieurs classes dans le classement « humain », ce qui n'est pas possible dans le classement Alhena.

Étape 2

Dans un deuxième temps, nous avons montré la galaxie obtenue avec Alhena à notre lectrice de l'étape 1 et nous lui avons demandé d'analyser les constellations locales et d'essayer de les nommer. Nous souhaitions vérifier si la galaxie avait une cohérence. Le tableau V.6 présente l'interprétation de la galaxie.

Cette étape nous a permis de voir que :
- les constellations locales correspondaient, assez souvent, à des regroupements nommables,
- la mesure de voisinage ne permettait pas de rapprocher des textes d'un même genre (par exemple des textes poétiques, des textes jeux de mots, ...),
- en revanche, la constellation peut servir de base à un classement réalisé par une ou plusieurs personnes. Cela se traduit par un gain de temps,
- Alhena a classé des textes mauvais syntaxiquement et sémantiquement.

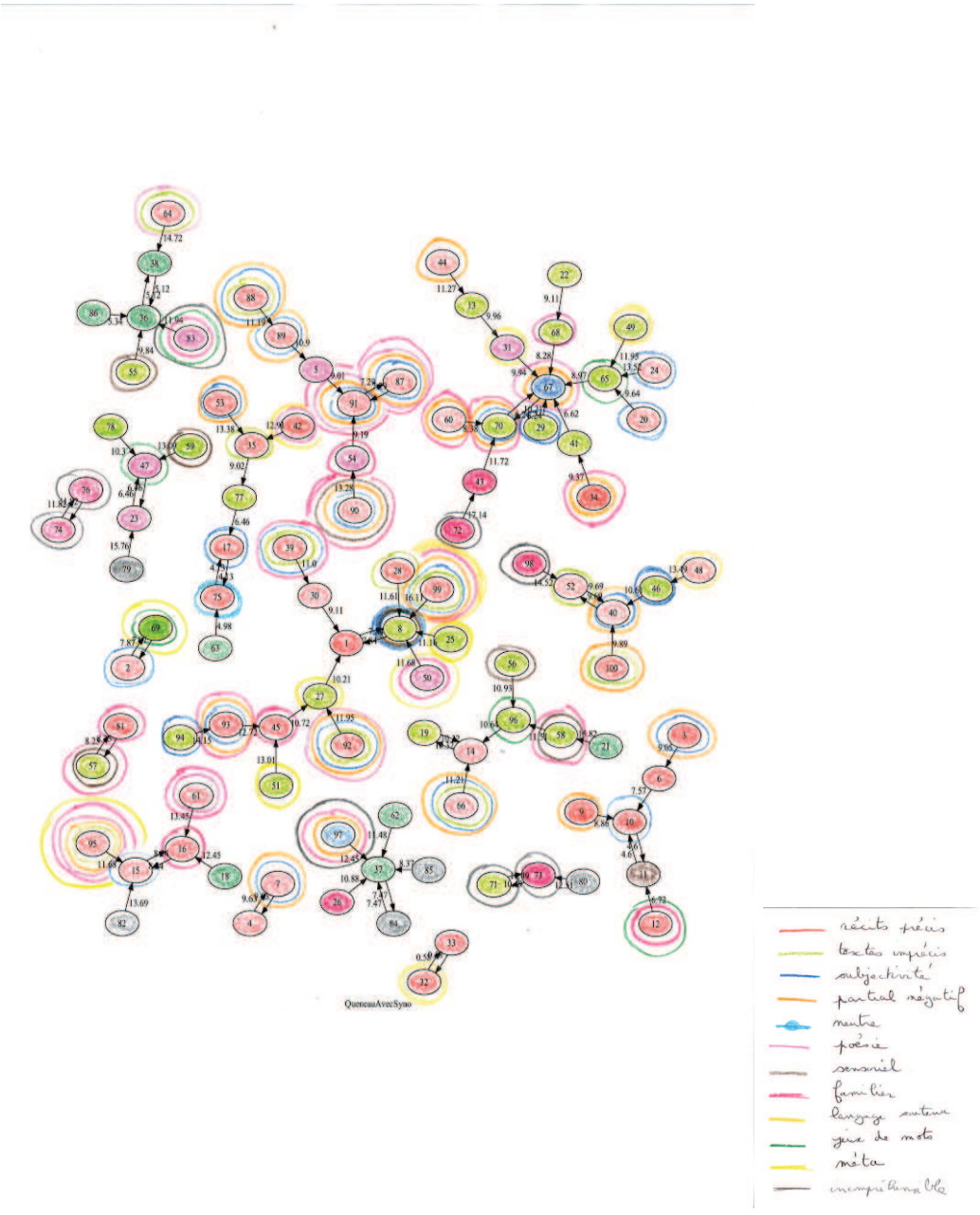


Figure V.9 – Galaxie obtenue avec Alhena sur la base des textes de Queneau

Numéro de la constellation locale	Interprétation
0	Difficile à nommer , dominante subjective
1	Récit, mêmes informations données
2	Pas de cohérence
3	CL trop diverse
4	Précision, objectif, mathématiques
5	Pas de cohérence
6	« Je », subjectif, informatif
7	Expression très proche
8	Répétition de consonne
9	Décalage avec le français « normal »
10	Passé
11	Textes incompréhensibles
12	Même degré d'information avec un jeune homme bête
13	Règnes animal et végétal
14	Vocabulaire du corps
15	Regroupement cohérent
16	Textes incompréhensibles

Tableau V.6 – Interprétation de la galaxie Queneau

4 Autres applications

4.1 Psychologie : Catégorisation de différentes situations de mémoire

Fanny Vallet et Olivier Desrichard s'intéressent au Sentiment d'Auto-efficacité Mnésique (SAM) qui est défini comme étant « *les croyances en ses propres capacités à utiliser sa mémoire efficacement dans des situations variées* » ([HHD89] cité par Vallet [Val12]). Pour connaitre l'évaluation que font les gens de leur mémoire, il est classique en psychologie d'utiliser des questionnaires élaborés par des experts de la mémoire. Or il n'est pas certain que les personnes naïves (i.e. non expertes) conçoivent la mémoire de la même façon. Donnons un exemple caricatural pour bien comprendre :

un des questionnaires de SAM existant est « avez-vous une bonne mémoire ? »

les experts ont des définitions précises de ce qu'est la mémoire. Mais qu'en est-il des personnes naïves ?

si un expert pense que la mémoire, c'est retenir des séries de chiffres et qu'une personne pense que la mémoire, c'est se souvenir de son enfance, ils ne parlent pas de la même chose et cela entraînera obligatoirement des différences de réponses à la question.

L'idée principale est de dire que le SAM est influencé par ce que les gens pensent de la mémoire. Fanny Vallet et Olivier Desrichard s'intéressent à ce que les personnes naïves pensent des domaines de mémoire (c'est à dire aux théories naïves des domaines de mémoire), ce qu'aucune étude n'avait fait auparavant. Pour cela, ils ont procédé à :

- des entretiens pour que des personnes naïves donnent des exemples de mémoire. Ces exemples ont été triés et sélectionnés, puis mis sous forme de cartes (ce sont donc les items).
- à ces exemples, ont été ajoutés des exemples d'items de questionnaires de SAM.
- Etude 1 : des participants de psychologie triaient les cartes en 2 tas : mémoire vs. non mémoire, de manière à ce que soient sélectionnés uniquement les items jugés consensuellement comme faisant appel à la mémoire. Ils devaient trier les items en tas, « en fonction de ce qui se ressemble du point de vue de la

mémoire ».

– Etude 2 : On a réduit le nombre de cartes et les participants devaient seulement trier les cartes en tas. Cette étude a été faite avec des personnes jeunes (sur lesquelles portent notre collaboration) et sur des personnes âgées.

Les participants « jeunes » de l'étude 2 ont trié 125 items de situations comme par exemple :

– « se souvenir de moments passés en famille »,
– « se souvenir du prix des choses »,
– « se souvenir des endroits où on est allé ».

Pour chaque couple d'items, le nombre de participants ayant rapproché ces 2 items a été comptabilisé. Ensuite une matrice a été construite : à la ligne i et à la colonne j se trouve le nombre de participants ayant associé l'item i et l'item j. Sur la diagonale, on trouve le nombre de participants car un item est par défaut rapproché de lui-même. Sur cette matrice, Fanny Vallet et Olivier Desrichard ont réalisé une classification hiérarchique basée sur une distance euclidienne. A partir de ce dendrogramme et

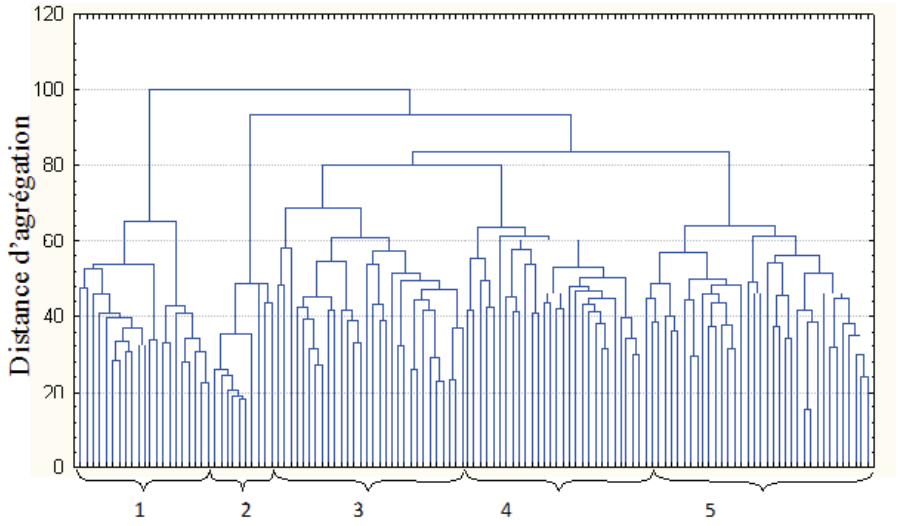


Figure V.10 – Dendrogramme lié aux situations de mémoire classées par les participants

de la classification, les chercheurs ont identifié 5 classes de situations de mémoire[7] (voir tableau V.7).

7. La labellisation des clusters est issue du consensus de 4 juges indépendants

Numéro	Nom de la classe	Sous-classes	
		Numéro	Nom de la sous-classe
1	Manquement de mémoire	1	Manquements de mémoire, général
		2	Manquements de mémoire concernant les personnes
2	Mémoriser un ensemble de choses et les rappeler ultérieurement		
3	Avoir et utiliser des connaissances	1	Mémoire des sens
		2	Se souvenir de choses apprises incidemment
		3	Récupérer des connaissances
4	Se souvenir de choses utiles au quotidien	1	Mémoire procédurale
		2	Un seul item
		3	Se souvenir de choses utiles pour l'organisation et la planification de la vie quotidienne
5	Mémoire autobiographique et émotionnelle	1	Se souvenir de détails
		2	Se souvenir d'évènements émotionnellement marqués
		3	Se souvenir d'évènements importants et marquants

Tableau V.7 – Classes identifiées par les chercheurs

L'objectif principal de notre collaboration était de comparer les résultats du classement « humain » et du classement « Alhena ». Fanny Vallet et Olivier Desrichard souhaitaient pouvoir répondre à une critique récurrente dans leur domaine : les regroupements effectués par les participants pourraient être dûs uniquement aux formulations des items. Ils pensent que les théories naïves qu'ont les participants sur la mémoire ont une influence sur le tri.
De notre côté, nous avons utilisé Alhena sur les 125 items de situations de mémoire (textes très courts, composés de quelques mots). Nous avons calculé la matrice des

mesures de voisinage entre chacun des textes. A partir de cette matrice, une analyse, semblable à l'analyse menée sur la matrice liée aux participants, a été menée et la figure V.11 présente le dendogramme résultat.

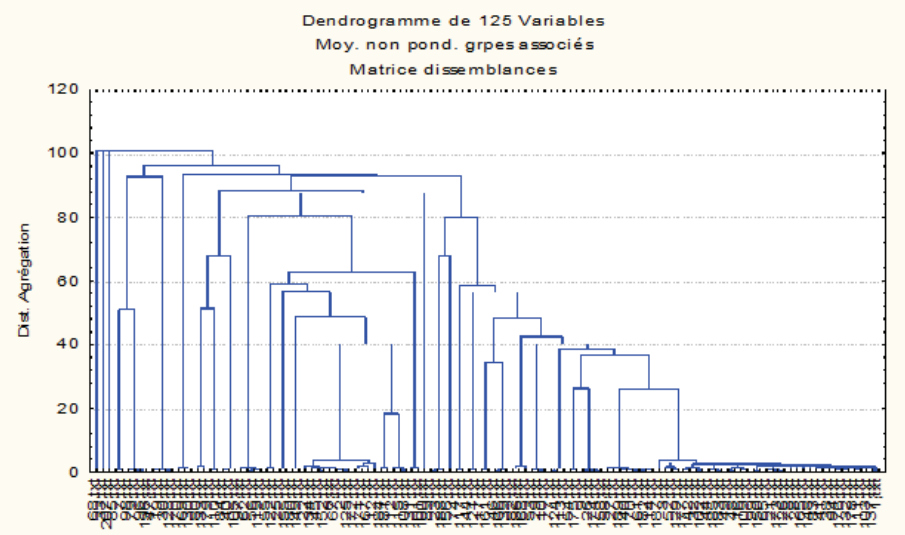


Figure V.11 – Dendrogramme lié à la matrice issue d'Alhena

A partir de la matrice créée par Fanny Vallet et Olivier Desrichard, nous avons tracé un graphe de manière analogue à la construction de graphe dans Alhena. Pour chaque item i , nous avons regardé l'item j que les participants avaient le plus souvent rapproché de l'item i. Dans le graphe, nous avons donc tracé une flèche de l'item i vers l'item j. La figure V.12 présente le graphe obtenu.
Nous avons créé une base de textes contenant l'ensmeble des items de mémoire (1 texte = 1 item). A l'aide d'Alhena, nous avons obtenu la galaxie présentée à la figure V.13

On note tout de suite que les deux dendogrammes et les deux galaxies sont très différents. Contrairement aux classes et constellations locales obtenues par rapport au classement « humain », les classes et les constellations locales obtenues sur la matrice Alhena sont difficilement nommables : en effet les textes sont très courts et contrairement aux expériences sur des textes longs où Alhena reconstruit implicitement l'univers sémantique, ici cela ne marche pas. De plus, les constellations locales du graphe obtenues à partir de la matrice des participants ont pu être nommées (voir tableau V.8). L'analyse par cluster et le graphe obtenu sur la matrice des

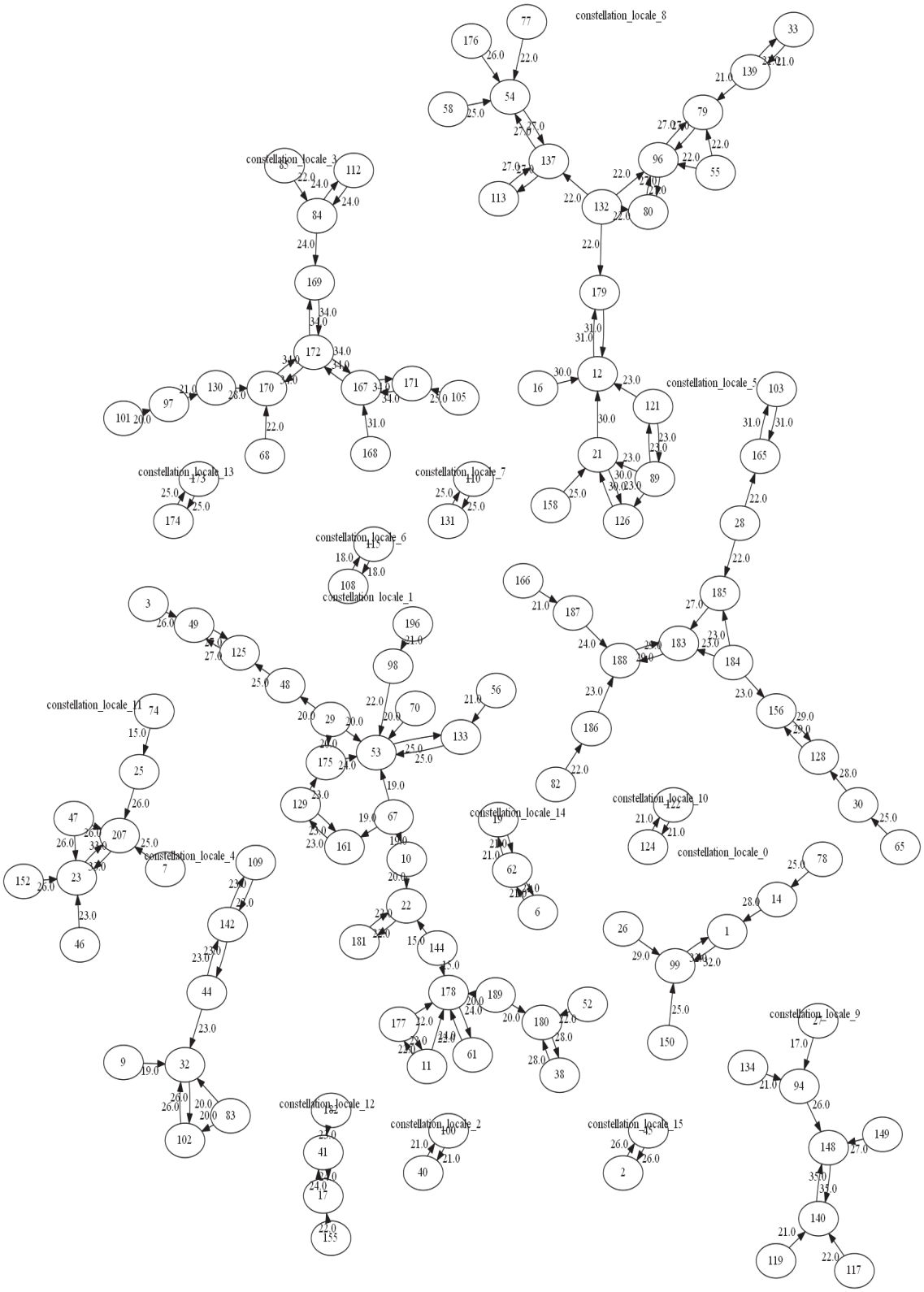


Figure V.12 – Graphe obtenu à partir de la matrice provenant des jugements des participants

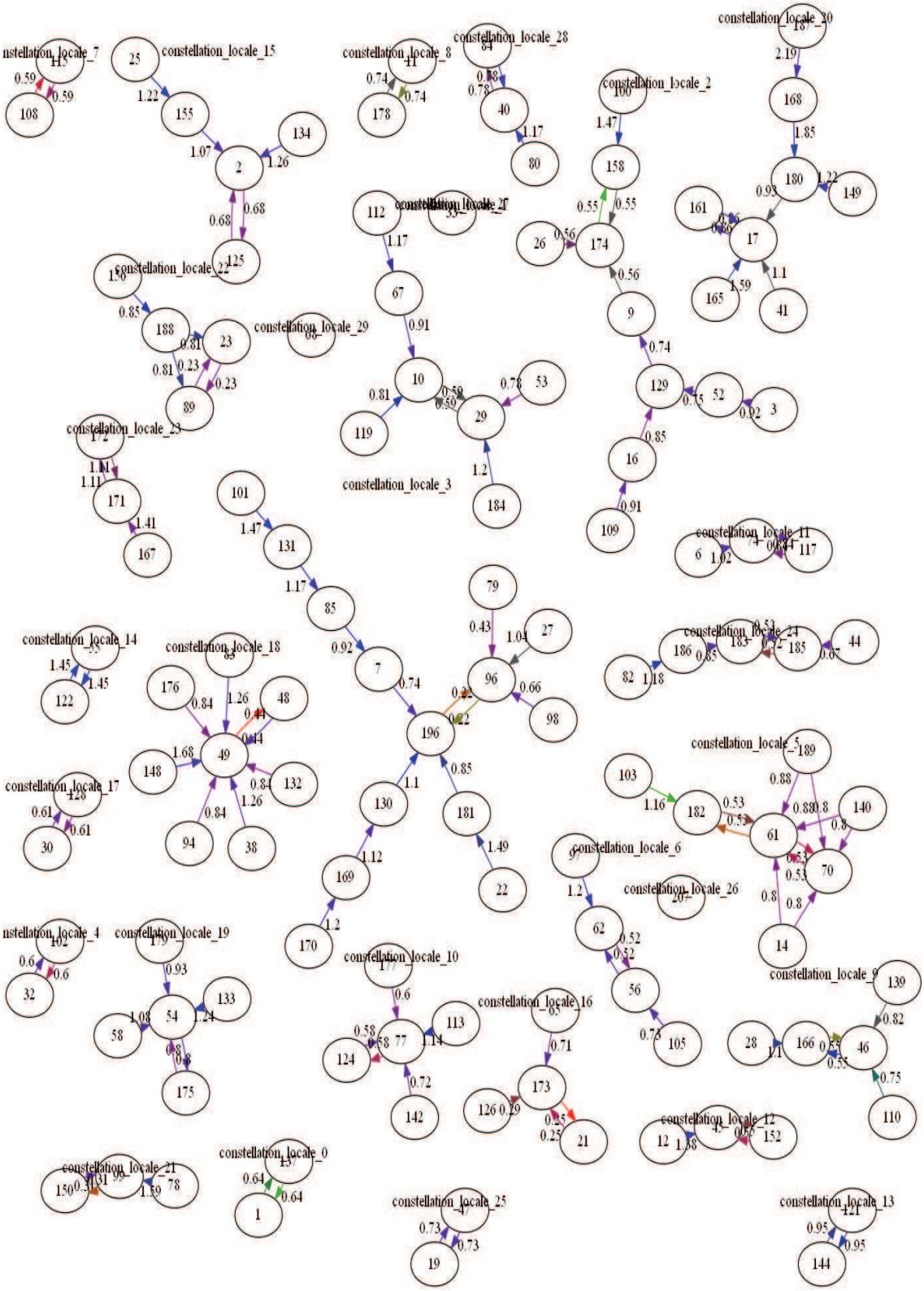


Figure V.13 – Galaxie obtenue sur les items des situations de mémoire

Numéro de la constellation locale	Nom
0	Mémoire autobiographique enfance/important
1	Se souvenir des choses utiles au quotidien
2	Mémoire procédurale
3	Mémoriser un ensemble de choses et se rappeler ultérieurement
4	Mémoire autobiographique lointaine
5	Récupérer des connaissances
6	Chanson/Sens
7	Retrouver un élément lointain
8	Manquements de mémoire
9	Mémoire autobiographique émotionnelle
10	Itinéraires
11	Connaissances
12	Connaissances retrouvées incidemment
13	
14	Mémoire implicite/automatique
15	Personnes

Tableau V.8 – Constellation locales à partir de la matrice des participants

participants donnent des résultats cohérents : à la seule différence que le graphe ne donne pas une hiérarchie.
En conclusion, cette collaboration a pu mettre en évidence que les participants n'ont donc pas seulement classé les situations de mémoire par rapport à leur formulation et plus précisément par rapport au vocabulaire utilisé mais par rapport à des théories naïves. Ce résultat était souhaité par les psychologues.

4.2 Informatique : Profil usager dans un contexte de collaboration

4.2.1 Présentation

Des chercheurs en informatique, Jean-Charles Marty et Thibault Caron, ont mené une expérience sur le travail collaboratif à travers un jeu [BMC11]. Ils ont développé un jeu de rôle « Learning Aventure » qui représente un environnement en 3D (voir figure V.14).

Pendant leurs activités d'apprentissage, les joueurs réalisent des quêtes de recherche

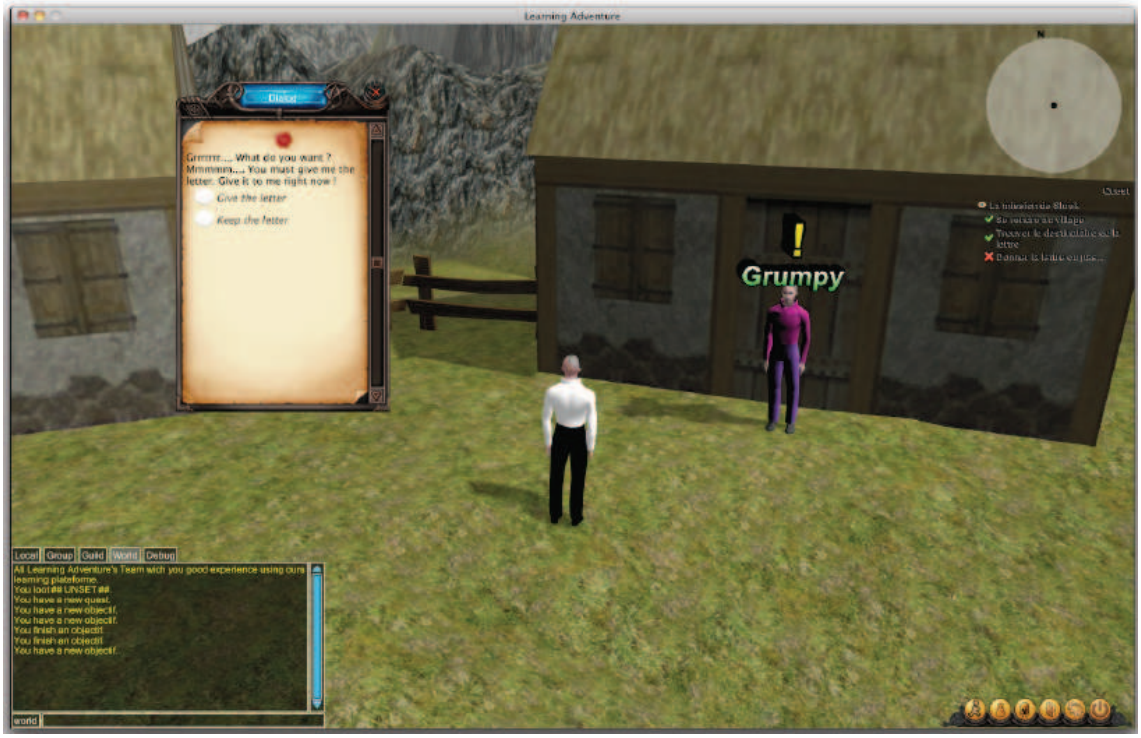


Figure V.14 – Capture d'écran du jeu Learning Aventure [BMC11]

de la connaissance, seuls ou en groupe, ce qui nécessite des dispositifs de collaboration au sein du jeu. Pour communiquer, les joueurs peuvent utiliser un outil de tchat disponible dans le jeu (coin inférieur gauche de la figure 1). D'autres outils (comme un mur à post it) peuvent être utilisés pour la collaboration. Dans cette expérimentation, les joueurs étaient des étudiants de première année d'IUT. Ces étudiants devaient réaliser deux tâches :

– première tâche : utiliser un outil collaboratif pour concevoir un jeu à partir de briques élémentaires. Les étudiants devaient découvrir certaines de ces briques élémentaires dans le monde (les briques sont directement issues de la classification des actions élémentaires dans le jeu (les briques sont par exemple la brique « tirer » ou la brique « éviter »...). Ensuite, les étudiants mettent en commun les briques qu'ils ont trouvées et utilisent un outil collaboratif pour écrire un plan descriptif de leur jeu.

– deuxième tâche : utiliser la méthode SCRUM [8] (méthode agile [9] dédiée à la gestion de projets) pour réaliser la liste des premières fonctionnalités à réaliser (« backlog produit ») pour l'implémentation de leur jeu.

Tout au long de l'expérimentation, des traces d'activités des étudiants sont enregistrées : par exemple leur communication dans le tchat, le parcours dans le monde... Certaines de ces traces sont utilisées au cours de l'expérimentation pour gérer les étudiants : voir l'avancement des étudiants sur les tâches, les étudiants en difficulté, quel étudiant utilise le plus l'outil collaboratif ... Le travail collaboratif permet également de voir le développement de sous-groupes et leur évolution dans les activités proposées.

Jean-Charles Marty et Thibault Caron étaient intéressés par le prototype Alhena pour utiliser une partie des traces collectées qui ne sont pour le moment pas exploitées. Ils s'intéressent aux rôles joués par les différents joueurs pendant la session : qui est leader ? Qui organise le travail ? etc. La figure V.15 présente des indicateurs sur les actions réalisées par un joueur : ainsi son profil peut être déterminé. Alhena devrait pouvoir aider à déterminer d'autres caractéristiques des joueurs : les joueurs partageant un vocabulaire proche.

4.2.2 Résultats

Dans un premier, nous avons créé un fichier texte pour chaque participant (étudiants et enseignant). Ce fichier contenait tout ce que le participant avait écrit dans le tchat. Notre base de textes contenait donc 16 fichiers. La figure V.16 présente les résultats obtenus sur cette base.

8. « La méthode s'appuie sur le découpage d'un projet en incréments, nommés "sprint". Ces étapes durent entre quelques heures et un mois (avec une préférence pour deux semaines). Chaque sprint commence par une estimation suivie d'une planification opérationnelle, au cours de laquelle l'équipe définit l'objectif du sprint. Le sprint se termine par une démonstration de ce qui a été réellement achevé. L'équipe analyse alors ce qui s'est passé durant ce sprint, afin de s'améliorer pour le prochain (rétrospective). »(source Wikipédia)

9. « Les méthodes agiles sont des groupes de pratiques pouvant s'appliquer à divers types de projets, mais se limitant plutôt actuellement aux projets de développement en informatique (conception de logiciel). Les méthodes agiles se veulent plus pragmatiques que les méthodes traditionnelles. Elles impliquent au maximum le demandeur (client) et permettent une grande réactivité à ses demandes. Elles visent la satisfaction réelle du besoin du client et non les termes d'un contrat de développement. »(source Wikipédia)

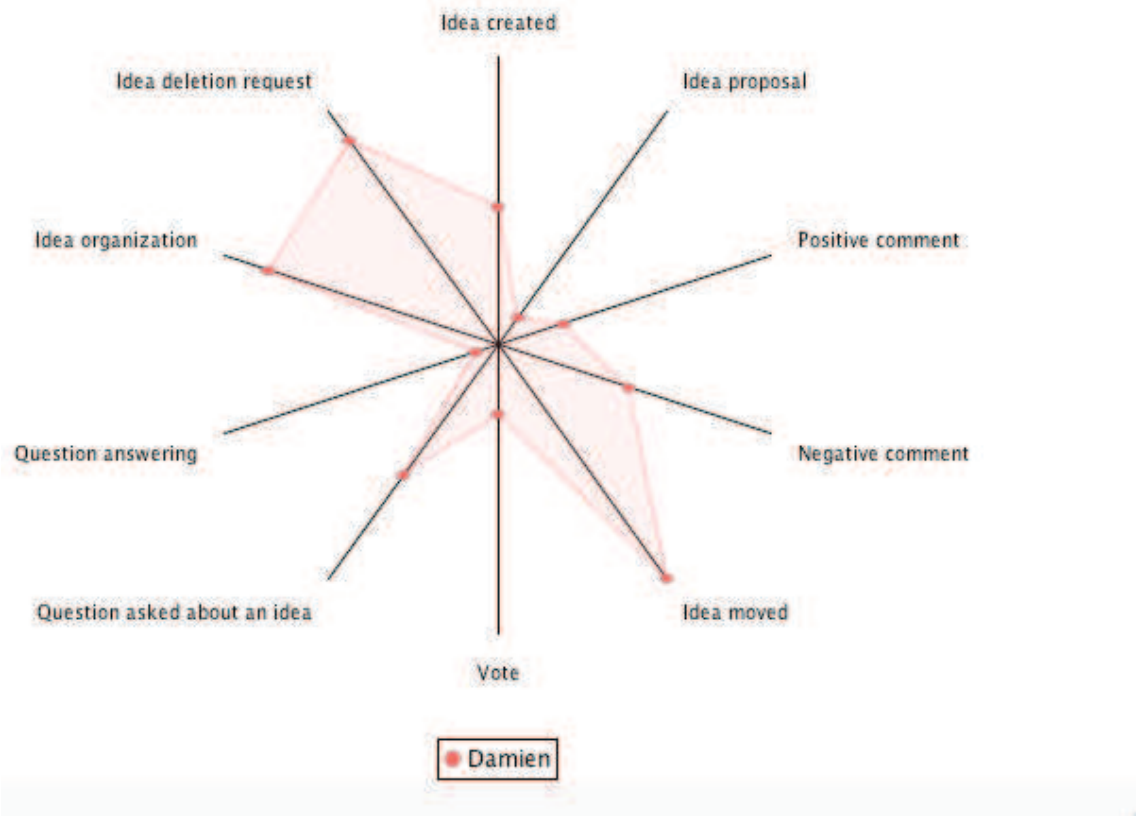


Figure V.15 – Indicateurs sur les actions d'un joueur

La galaxie obtenue nous montre les points suivants :

- la personne qui a lancé le jeu a été isolée des joueurs. Son discours dans le tchat constitue à lui seul une constellation locale,
- l'analyse des deux autres constellations locales montre que les regroupements correspondent aux deux phases du jeu. La constellation locale 0 regroupe un ensemble de joueurs qui ont collaboré et le rapprochement s'est fait sur le vocabulaire lié à la première tâche. Dans la constellation locale 1, c'est sur le vocabulaire lié à la deuxième tâche que le regroupement a été réalisé. Ici, on remarque que le premier groupe a surtout échangé sur la première tâche à réaliser alors que le second a davantage discuté lors de la deuxième tâche.

Le tchat étant assez court (peu d'échanges entre les joueurs) dans cette expérience, nous n'avons pas pu dégager de profil utilisateur mais les premières observations

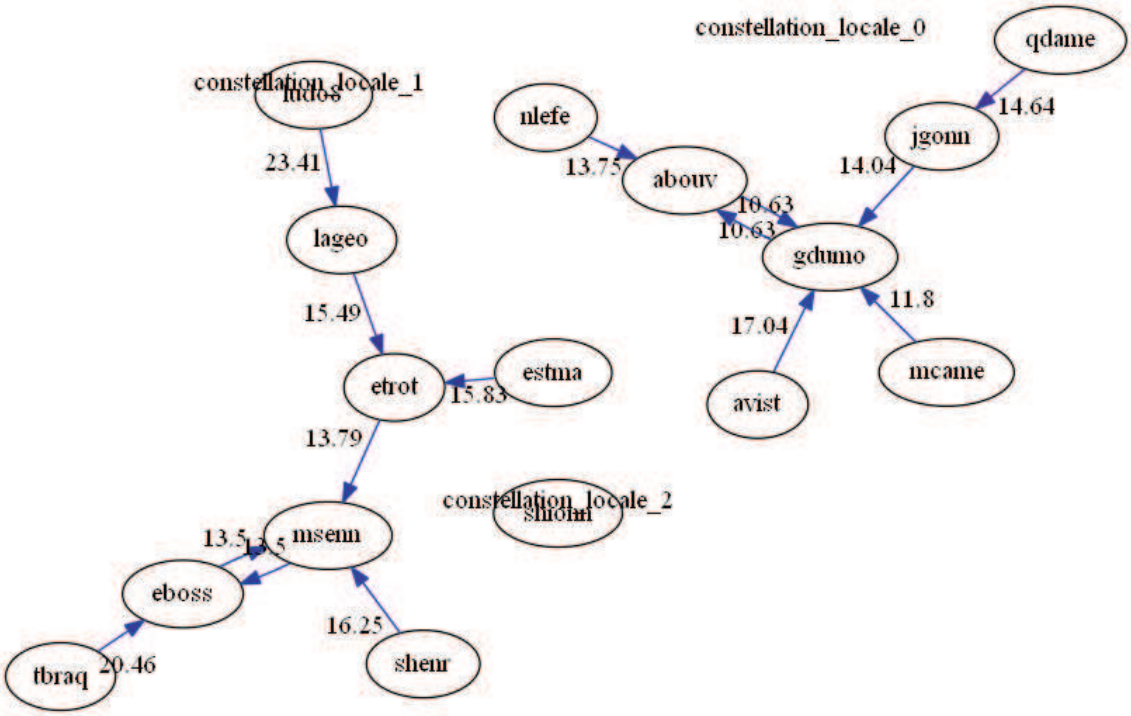


Figure V.16 – Galaxie obtenue sur le tchat du jeu Learning Aventure

sont encourageantes. Par conséquent, nous allons mener d'autres expériences sur des tchats plus conséquents.

Références bibliographiques

[BMC11] M. Bodin, J.-C. Marty et T. Carron : Specifying collaborative tools in game-based learning environments : Clues from the trenches. *In Proc. of the 5th European Conference on Games Based Learning (ECGBL), Athens, Grece, 2011*, pages pp46–56, 2011. 172, 173

[HHD89] C. Hertzog, D.F. Hultsch et R.A. Dixon : Evidence for the convergent validity of two self-report metamemory questionnaires. *Developmental Psychology*, 25(5):687, 1989. 166

[MB04] A. Moschitti et R. Basili : Complex linguistic features for text classification : A comprehensive study. *Advances in Information Retrieval*, pages 181–196, 2004. 152

[Que47] R. Queneau : *Exercices de style*. Gallimard, 1947. 141, 159

[Seb02] F. Sebastiani : Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47, 2002. 152

[Val12] F. Vallet : *Comment évalue-t-on l'efficience de notre mémoire : Le rôle des attributions causales et des théories naïves*. Thèse de doctorat, Université de Grneoble, 2012. 166

[Yan99] Y. Yang : An evaluation of statistical approaches to text categorization. *Information retrieval*, 1(1):69–90, 1999. 152

Conclusion générale et perspectives

L'objectif de cette thèse était de répondre à une problématique de la veille stratégique : créer un outil permettant à l'animateur de sélectionner de manière simple et rapide des informations voisines. Nous avons tout d'abord défini le concept d'information voisine. Ensuite, nous avons proposé une mesure évaluant la proximité entre deux textes en s'appuyant sur les mots communs, les synonymes et les mots cooccurrents de ces textes. Cette mesure a été implémentée dans un outil que nous avons nommé Alhena. Enfin, cet outil a été testé en veille stratégique sur le domaine de la valorisation du CO_2. Nous avons également procédé à une validation sur une base de textes classée et nous avons utilisé Alhena dans trois projets différents : un en littérature, un en psychologie et un autre en informatique.

Nous avons sommes partis d'un problème concret pour proposer une mesure de voisinage entre textes et un outil informatique permettant :

- une navigation facile dans un graphe appelé galaxie,
- un regroupement des textes sous forme de constellations locales,
- un étiquetage lié aux constellations locales et aux textes permettant une lecture plus rapide.

Nous avons montré également que les galaxies avaient les propriétés suivantes :

- pour chaque constellation locale, au moins un nucléus réalisant la mesure minimale,
- le long des bras des constellations locales, la mesure diminue en se rapprochant du nucléus,

- le long des bras, le nombre de mots composant les textes est de plus en plus petit en se rapprochant du nucléus : le bras d'une constellation locale réalise une analyse des documents (on va du général vers le particulier) ; cette approche est différente de l'approche « cosinus » qui propose une synthèse des documents composants le bras de la constellation (on va du particulier vers le général).
- la possibilité de détecter les doublons et les textes inclus dans d'autres textes.

Nous avons montré que notre mesure de voisinage proposait un classement ayant de bons scores de rappel et de précisions sur la base classée Reuters.

Dans le domaine de la veille stratégique, l'objectif visé était d'apporter des aides pour surmonter la paralysie occasionnée par les gros volumes de données numériques et, par voie de conséquence, pour diminuer les coûts administratifs liés à la veille. L'expérimentation effectuée à l'aide d'Alhena a permis de constater :

- un gain considérable de temps pour l'exploration des FULL text tirés des sources Internet (réduction d'une vingtaine d'heures à deux heures environ, dans le cas de l'expérimentation effectuée) ;
- une facilitation pour rechercher des informations de nature à se fiabiliser mutuellement ;
- une facilitation pour suggérer des liens entre des informations (construction des puzzles utilisant les "pièces" d'information que sont les signaux faibles) de la part des participants au groupe de travail chargé d'interpréter les signaux faibles (lien d'incohérence entre deux informations ; lien de complémentarité ; lien de causalité entre deux informations, etc.)
- le prototype ne nécessite pas de fichiers textes structurés (par exemple avec des balises pour reconnaître le titre, l'auteur, la date, ...). Cette spécificité nous permet de comparer des textes issus de sources différentes.
- Alhena est applicable pour n'importe quelle langue, si l'on dispose d'un lemmatiseur pour celle-ci.

Nous avons pu également tester notre mesure de voisinage dans d'autres domaines que la veille stratégique :

- en littérature : bien que certains textes de Queneau soit syntaxiquement mauvais, nous avons pu proposer une classification,
- en psychologie : nous avons pu comparer notre classement à un classement « humain » de textes très courts (quelques mots),
- en informatique : Alhena a proposé un regroupement pour des textes issus de tchats, textes où l'orthographe est souvent incorrecte et dans lesquels on trouve un grand nombre de mots inconnus des dictionnaires.

Notre prototype a pu classer des textes de différentes natures (littéraires, journalistiques, tchats, ...) et a permis un gain de temps de lecture.

Notre prototype présente quelques faiblesses :

- Alhena est, en l'état actuel du prototype, incapable de traiter des masses colossales de textes (par exemple le web) dans un temps raisonnable : pour le cas CO_2, il propose une représentation graphique sur 300 textes au bout d'une heure,
- le prototype se heurte au problème d'encodage des fichiers : Alhena nécessite des fichiers txt au format accepté par le lemmatiseur,
- Alhena est un outil d'aide à la lecture : il propose à un moment donné une classification des textes parmi toutes celles possibles. La classification proposée n'est sans doute pas la « proposition idéale »,
- le prototype ne propose pas d'effectuer des recherches basées sur la sémantique ou la syntaxe : néanmoins Alhena ne s'attachant pas à la syntaxe ou à l'orthographe regroupe même des textes qui comportent des erreurs de syntaxe ou qui sont sémantiquement mauvais,
- le classement est plus difficile dans le cas de textes très courts (une phrase de quelques mots),
- Alhena n'est pas, pour l'instant, un outil multilingue : il travaille sur une seule langue à la fois. Cependant, Alhena peut traiter des bases de textes en français, anglais et italien.
- Alhena n'est pas pour l'instant un outil « clef en main ».

Nous envisageons donc :

- d'étudier les autres « minimuns » : nous nous sommes intéressés seulement aux textes réalisant la mesure minimale. Or la matrice des mesures de voisinage est riche et nous devons l'étudier.
- d'optimiser le code pour traiter, dans un avenir proche, plus de données dans un temps raisonnable,
- un développement informatique pour transformer Alhena en un outil « clef en main » sera réalisé.

De plus de nouvelles expérimentations vont être menées :

- en veille stratégique sur de nouveaux domaines,
- en veille stratégique en se combinant avec l'application Aproxima [10] ,
- en littérature : par exemple, les Exercices de style ont été traduits dans de nombreuses langues notamment en italien par Umberto Eco et nous souhaitons faire une comparaison entre le classement des textes en français et celui des textes en italien. Nous voulons étudier le processus de traduction.
- en informatique : des tchats d'expériences plus conséquents vont être analysés à l'aide d'Alhena,
- proposer un classement des dépêches AFP ou Reuters en temps réel,
- tracer l'information sur internet et ainsi suivre l'évolution et l'appropriation par différentes communautés d'une même information,
- en sciences politiques : effectuer un classement des discours collectés lors d'une campagne pour des élections (par exemple les élections présidentielles),
- d'autres applications sont possibles : classement de documents d'entreprise, classement des CV pour des directeurs de ressources humaines, ...

10. Approxima, développé par Alex Buitrago, est un outil de collecte et de sélection des diverses informations numériques d'Internet ; il offre des fonctionnalités intéressantes de filtre et d'analyse d'informations brutes afin d'aider l'utilisateur à identifier les quelques informations susceptibles d'avoir un intérêt pour la phase d'exploitation des informations de veille et notamment des signaux faibles.

Références bibliographiques

Chapitre

[age] AgentIntelligent.com - veille strategique - définition et objectifs. http ://www.agentintelligent.com/veille/veille_strategique.html. 2

[CKI07] E. Campoy, M. Kalika et H. Isaac : Surcharge informationnelle, urgence et tic : l'effet temporel des technologies de l'information. *Management et Avenir*, 13:153, 2007. 1

[Gui12] H. Guillaud : Notre surcharge informationnelle en perspective. http ://www.internetactu.net/2012/02/29/lift12-notre-surcharge-informationnelle-en-perspective/, 2012. 1

[Les03] H. Lesca : *Veille stratégique : la méthode L.E.SCAnning*. Gestion en liberté, ISSN 1625-3132. EEd. EMS, Colombelles, 2003. 2

[SJ12] A. Saint-Jude : From gutenberg to zuckerberg, information overload and social networks, 2012. 1

[SR10] C. Sauvajol-Rialland : La surcharge informationnelle dans l'organisation : les cadres au bord de la « crise de nerf ». *Le magazine de la communication de crise et sensible*, 19, 2010. 2

Chapitre I

[AC94] E. Auster et C.W. Choo : How senior managers acquire and use information in environmental scanning. *Information Processing & Management*, 30(5):607–618, 1994. 18

[AFN98] AFNOR : Prestations de veille et prestations de mise en place d'un système de veille. Rapport technique, AFNOR, 1998. 8, 19, 21

[Agu67] F.J. Aguilar : *Scanning the business environment.* Macmillan New York, 1967. 17, 18

[Ans75] H. I. Ansoff : Managing strategic surprise by response to weak signals. *California Management Review*, 18(2):21–33, 1975. 11

[Arg96] C. Argyris : Actionable knowledge : Design causality in the service of consequential theory. *The Journal of Applied Behavioral Science*, 32(4): 390, 1996. 8

[Bon10] J. Bondu : Benchmarking des plateformes de veille choisir son outil. Rapport technique, Inter-Ligere et SerdaLab, 2010. 39

[BR09] D. Bawden et L. Robinson : The dark side of information : overload, anxiety and other paradoxes and pathologies. *Journal of Information Science*, 35(2):180–191, avril 2009. 41

[CA93] C.W. Choo et E. Auster : Environmental scanning : acquisition and use of information by managers. *Annual Review of Information Science and Technology*, 28:279–314, 1993. 18

[CFF03] M.-L. Caron-Fasan et A. Farastier : *Veille stratégique et gestion des connaissances*, pages 237–266. 2003. 29, 30, 31, 32, 33

[Cul83] M.J. Culnan : Environmental scanning : The effects of task complexity and source accessibility on information gathering behavior. *Decision Sciences*, 14(2):194–206, 1983. 18

[Dav00] A. David : La recherche intervention, un cadre général pour les sciences de gestion. *In IX conférence internationale de management stratégique, Montpellier*, pages 24–26, 2000. 8

[Dou05] H. Dou : Veille technologique en formulation. *Techniques de l'ingénieur. Génie des procédés*, (J2260), 2005. 16

[DSP88] R.L. Daft, J. Sormunen et D. Parks : Chief executive scanning, environmental characteristics, and company performance : An empirical study. *Strategic Management Journal*, 9(2):123–139, 1988. 18

[DW84] R.L. Daft et K.E. Weick : Toward a model of organizations as interpretation systems. *Academy of management review*, pages 284–295, 1984. 18, 28

[Ele97] D.S. Elenkov : Strategic uncertainty and environmental scanning : the case for institutional influences on scanning behavior. *Strategic Management Journal*, 18(4):287–302, 1997. 18

[EM04] M.J. Eppler et J. Mengis : The concept of information overload : A review of literature from organization science, accounting, marketing, mis, and related disciplines. *The information society*, 20(5):325–344, 2004. 9, 23, 36

[FB03] A. Farastier et B. Ballaz : Le management des connaissances et la fonction achats : la dimension interorganisationnelle et le rôle clé du système d'information. *Cahier de recherche du CERAG*, 2003. 30

[FHH84] J.L. Farh, R.C. Hoffman et W.H. Hegarty : Assessing environmental scanning at the subunit level : A multitrait-multimethod analysis. *Decision Sciences*, 15(2):197–220, 1984. 18

[FI04] L. Favier et M. Ihadjadene : *Les outils de veille et d'intelligence économique*, chapitre 10. Hermes Science Publications, mar 2004. 19, 40

[Fra05] C. François : L'analyse de l'information proposée par les outils de veille. *Regards sur l'IE : le magazine de l'intelligence économique*, 11:60–63, Oct 2005. 37, 39

[Gho88] S. Ghoshal : Environmental scanning in korean firms : organizational isomorphism in action. *Journal of International Business Studies*, pages 69–86, 1988. 18

[Gue09] M. Guechtouli : Comment organiser son système de veille stratégique ? *In Symposium ATELIS*, 2009. 19, 20

[Ham81] D.C. Hambrick : Specialization of environmental scanning activities among upper level executives. *Journal of Management Studies*, 18(3): 299–320, 1981. 18

[Ham82] D.C. Hambrick : Environmental scanning and organizational strategy. *Strategic Management Journal*, 3(2):159–174, 1982. 18

[Hem09] P. Hemp : Death by information overload. *Harvard Business Review*, 87(9):82–89, 121, septembre 2009. PMID : 19736853. 37

[Huf90] A.S. Huff : *Mapping strategic thought*. John Wiley & Sons, 1990. 29

[IFC03] M. Ihadjadene, L. Favier et S. Chaudiron : L'intelligence économique sur internet : évaluation des pratiques en france. *Présenté à Intelligence Economique : Recherches et Applications*, 2003. 39

[JMFL06] R. Janissek-Muniz, H. Freitas et H. Lesca : Veille anticipative stratégique, intelligence collective (vas-ic) : usage innovant du site web pour la provocation d'informations d'origine terrain. *La Revue des Sciences de Gestion, Direction et Gestion*, (218):19–30, 2006. 17

[KC05] S. Kamoun-Chouk : *Veille Anticipative Stratégique : Processus d'Attention à l'Environnement Application à des PMI tunisiennes*. Thèse de doctorat, Université Pierre Mendès France (Grenoble), 2005. 11, 17, 18, 29

[KEN05] Y. KENGEN : Métiers de veille : survivre à la technologie. la veille, les outils froids et la norme afnor xp x50-053. *In Veille stratégique, scientifique et technologique*, 2005. 19

[LD10] H. Lesca et R. Dourai : Traque et remontée des informations de veille stratégique anticipative : une approche par la notion d'épanouissement de soi. *FACEF Pesquisa*, 7(2), 2010. 30

[LE86] R.T. Lenz et J.L. Engledow : Environmental analysis : The applicability of current theory. *Strategic Management Journal*, 7(4):329–346, 1986. 18

[Les86] H. Lesca : *Système d'information pour le management stratégique de l'entreprise : l'entreprise intelligente*. McGraw-Hill, 1986. 8

[Les92] H. Lesca : Le problème crucial de la veille stratégique : la construction du "puzzle". 1992. 25

[Les03a] H. Lesca : *Veille stratégique : la méthode L.E.SCAnning*. Gestion en liberté, ISSN 1625-3132. EEd. EMS, Colombelles, 2003. 9, 17, 19, 20, 21, 22, 23, 24, 28

[Les03b] N. Lesca : La veille stratégique : vers un système d'information pour le management stratégique des discontinuités. *In Présent et futurs des systèmes d'information*. PUG, aug 2003. 19

[Les08] M.-L. Lesca, N. et Caron-Fasan : Facteurs d'échec et d'abandon d'un projet de veille stratégique : retour d'expériences. *Systèmes d'Information et Management*, 13(3), 2008. 37

[Les10] H. Lesca : *Chimie durable et signaux faibles : le cas du CO2 vu comme une matière à valoriser*, chapitre 110. Traités IC2. Série Technologies et développement durable. Lavoisier, Paris : Hermès science publications, 2010. 34, 35

[LKC09] H. Lesca, S. Kriaa et A. Casagrande : Veille stratégique : Un facteur d'échec paradoxal largement avéré : la surinformation causée par l'internet. cas concrets, retours d'expérience et piste de solutions. *La revue des sciences de gestion*, (245-246):35–42, 2009. 36

[LL10] H. Lesca et E. Lesca : *Gestion de l'Information*. EMS, 2010. 10

[LL11] H. Lesca et N. Lesca : *Les signaux faibles et la veille anticipative pour les décideurs : Méthodes et applications*. Hermes Science Publications, mai 2011. 10, 11, 12, 14, 16, 21, 24, 25, 28, 34, 40

[ML10] S. Medhaffer et H. Lesca : *L'animation de la veille stratégique*. Hermes Science Publications, février 2010. 8, 24, 33

[MSJS00] R.C. May, W.H. Stewart Jr et R. Sweo : Environmental scanning behavior in a transitional economy : evidence from russia. *Academy of Management Journal*, pages 403–427, 2000. 18

[MW04] A.V. Medeiros Wanderley : *Conception et implantation d'un système d'intelligence compétitive dans une entreprise pétrolière dans un environnement de déréglementation*. Thèse de doctorat, Université Paul Cézanne (Aix-Marseille), 2004. 16

[NK98] I. Nonaka et N. Konno : The concept of" ba" : Building a foundation for knowledge creation. *California Management Review*, 40(3):40–54, 1998. 29, 31, 32

[Non91] I. Nonaka : The knowledge-creating company. *Harvard Business Review*, 69(6):96–104, 1991. 30

[NT95] I. Nonaka et H. Takeuchi : *The knowledge-creating company : How Japanese companies create the dynamics of innovation*. Oxford University Press, USA, 1995. 30

[NT03] I. Nonaka et R. Toyama : The knowledge-creating theory revisited : knowledge creation as a synthesizing process. *Knowledge Management Research & Practice*, 1(1):2–10, 2003. 28, 29, 32

[PS78] J. Pfeffer et G.R. Salancik : *The external control of organizations : A resource dependence perspective*. Harper and Row, 1978. 18

[Rab] F. Rabat : Petite contribution à une définition opératoire du concept d'information. http ://docsdocs.free.fr/spip.php ?article374. 9

[SA97] M. Salles et A.-M. Alquier : Réflexions méthodologiques pour la conception de systèmes d'intelligence économique de l'entreprise. *actes du Congrès international " le Génie Industriel dans un monde sans frontières*, pages 3–5, 1997. 19

[SFN88] L.R. Smeltzer, G.L. Fann et V.N. Nikolaisen : Environmental scanning practices in small business. *Journal of Small Business Management*, 26(3):55–62, 1988. 18

[SM99] O.O. Sawyerr et J.E. McGee : The impact of personal network characteristics on perceived environmental uncertainty : an examination of owners/managers of new high technology firms. *Frontiers of Entrepreneurship Research*, 1999. 18

[Tho03] J.D. Thompson : *Organizations in action : Social science bases of administrative theory*. Transaction Pub, 2003. 18

[Wei79] K.E. Weick : *The social psychology of organizing*, volume 2. Addison-Wesley, 1979. 29

[YLBH10] A. Yurchyshyna, M. Léonard et P. Brough-Heinzman : Towards a services-based approach for supporting idea development process. *In Proceedings of the 2010 Fifth International Conference on Internet and Web Applications and Services*, ICIW '10, pages 321–326, Washington, DC, USA, 2010. IEEE Computer Society. 8

Chapitre II

[Bal90] J.P. Balpe : *Hyperdocuments, hypertextes, hypermédias*. Eyrolles, 1990. 55

[Baz05] M. Baziz : *Indexation conceptuelle guidée par ontologie pour la recherche d'information*. Thèse de doctorat, [s.n.], [S.l.], 2005. 59, 65

[BCD08] N. Bonnel, M. Chevalier et B. Dousset : *Métaphores de visualisation des résultats de recherche d'information sur le web*. Lavoisier, 2008. 74, 75, 76

[Bes01] R. Besançon : *Intégration de connaissances syntaxiques et sémantiques dans les représentations vectorielles de textes*. Thèse de doctorat, Lausanne, 2001. 66

[Bis00] G. Bisson : La similarité : une notion symbolique/numérique. *Apprentissage symbolique-numérique (tome 2). Eds Moulet, Brito. Editions CEPADUES. pp. 169*, 201, 2000. 66

[BKL04] M.W. Bilotti, B. Katz et J. Lin : What works better for question answering : Stemming or morphological query expansion ? *In Proceedings of the Information Retrieval for Question Answering (IR4QA) Workshop at SIGIR 2004*, 2004. 58

[BKN04a] M. Boughanem, W. Kraaij et J.-N. Nie : *Modèles de langue pour la recherche d'information*, page 163–182. Lavoisier, 2004. 66

[BKN04b] M. Boughanem, W. Kraaij et J.Y. Nie : Modeles de langue pour la recherche d'information. *Les systemes de recherche d'informations*, pages 163–182, 2004. 71

[BLM03] J.-P. Barthélémy, X. Luong et S. Mellet : Prenons nos distances pour comparer des textes, les analyser et les représenter. *Corpus*, (2), 2003. 67

[Bon06] N. Bonnel : *Génération dynamique de présentations interactives en multimédia 3D, de données, pour les applications en ligne*. These, Université Rennes 1, décembre 2006. 73, 76, 77

[BP93] M. Bécue et R. Peiro : Les quasi-segments pour une classification automatique des réponses ouvertes. *Actes des 2ndes Journées Internationales d'analyse des données textuelles*, pages 310–325, 1993. 60

[Bru03] E. Brunet : Peut-on mesurer la distance entre deux textes ? *Corpus*, (2), 2003. 67

[BS08] M. Boughanem et J. Savoy : *Recherche d'information : état des lieux et perspectives*. Collection Recherche d'information et web, ISSN 1968-8008. Lavoisier, Paris : Hermès science publ., 2008. 51, 52, 59, 61, 62, 80

[BYRN99] R. Baeza-Yates et B. Ribeiro-Neto : *Modern information retrieval*. 1999. 49, 50, 54, 57

[CCT10] S.-S. Choi, S.-H. Cha et C.C. Tappert : A survey of binary similarity and distance measures. *Journal of Systemics, Cybernetics and Informatics*, 8(1):43–48, 2010. 69

[CDM07] Y. Champclaux, T. Dkaki et J. Mothe : Utilisation des similarités structurelles pour l'évaluation de la pertinence en recherche d'information. *In Colloque Veille Stratégique Scientifique et Technologique (VSST 2007), Marrakech (Maroc), 21/10/2007-25*, volume 10, page 2007, 2007. 72

[CDM10] Y. Champclaux, T. Dkaki et J. Mothe : An information retrieval models taxonomy based on an analogy between cognitive science and information retrieval. *In Actes colloque VSST'10*, 2010. 68, 72

[CG11] S. Clinchant et É. Gaussier : *Modèles probabilistes pour la recherche d'information*. Collection Recherche d'information et web, ISSN 1968-8008. Lavoisier, Paris : Hermès science publications, 2011. 53, 71

[Che02] M. Chevalier : *Interface adaptative pour l'aide à la recherche d'information sur le web*. These, Université Paul Sabatier - Toulouse III, Dec 2002. 74, 75

[CLS00] J. Cugini, S. Laskowski et M. Sebrechts : Design of 3-d visualization of search results- evolution and evaluation. *Visual data exploration and analysis VII*, pages 198–210, 2000. 74

[CV07] R. Cilibrasi et P. Vitanyi : The google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, 19(3):370–383, 2007. 66

[EGRCG+06] A. El Golli, F. Rossi, B. Conan-Guez, Y. Lechevallier *et al.* : Une adaptation des cartes auto-organisatrices pour des données décrites par un tableau de dissimilarités. *Revue de statistique appliquée*, 54(3):33–64, 2006. 71

[GY11] É. Gaussier et F. Yvon : *Modèles statistiques pour l'accès à l'information textuelle*. Collection Recherche d'information et web, ISSN 1968-8008. Lavoisier, Paris : Hermès science publications, 2011. 51

[Hei04] S. Heiden : Interface hypertextuelle à un espace de cooccurrences : implémentation dans weblex. *In* Anne Dister Gérard Purnelle, Cédrick Fairon, éditeur : *Le poids des mots*, volume 1, pages 577–588, Louvain-la-Neuve, Belgique, Mar 2004. Presses Universitaires de Louvain. 66

[Her06] N. Hernandez : *Ontologies de domaine pour la modélisation du contexte en recherche d'Information*. Thèse de doctorat, [s.n.], [S.l.], 2006. 59

[Iha04] M. Ihadjadene : *Les systèmes de recherche d'informations : modèles conceptuels*. Traité des sciences et techniques de l'information. Lavoisier, Paris : Hermès science publ., 2004. 52, 60

[IS07] F. Ibekwe-Sanjuan : *Fouille de texte*. Systèmes d'information et organisations documentaires. Hermès - Lavoisier, Mar 2007. 49, 50, 71

[IV98] N. Ide et J. Véronis : Introduction to the special issue on word sense disambiguation : the state of the art. *Comput. Linguist.*, 24:2–40, mars 1998. 65

[Joa98] T. Joachims : Text categorization with support vector machines : Learning with many relevant features. *Machine Learning : ECML-98*, pages 137–142, 1998. 71

[Joy09] J. Joyce : *Ulysses*. Echo Library, 2009. 62

[LC91] H. Le Crosnier : Une introduction à l'hypertexte. *BBF*, (4):280 – 294, 1991. 54

[Lem08] B. Lemaire : Limites de la lemmatisation pour l'extraction de significations. *In Actes des 9e Journées internationales d'Analyse Statistique des Données Textuelles (JADT'2008)*, pages 725–732, Lyon, France, 2008. 58

[LL03] C. Labbé et D. Labbé : La distance intertextuelle. *Corpus*, (2), 2003. 71

[Loi04] Y. Loiseau : *Recherche flexible d'information par filtrage flou qualitatif*. Thèse de doctorat, Université Paul Sabatier (Toulouse), 2004. 51, 52

[May05] D. Mayaffre : De la lexicométrie à la logométrie. *Astrolabe*, pages 1–11, 2005. 58

[MB09] S. Mellet et J.-P. Barthélemy : La topologie textuelle : légitimation d'une notion émergente. *Lexicometrica*, 7:http ://www.cavi.univ–paris3.fr/lexicometrica/numspeciaux/special9/mellet.pdf, 2009. 57

[MCA01] J. Mothe, C. Chrisment et J. Alaux : Visualisation globale de collections de documents sous forme d'hypercube. *Extraction des Connaissances et Apprentissage (ECA) journal*, 4:131–142, 2001. 76

[Mel01] S. Mellet : Lemmatisation et encodage grammatical : un luxe inutile ? *Lexicometrica*, (numéro spécial "Autour de la lemmatisation"), 2001. 58

[Mem00] D. Memmi : Le modèle vectoriel pour le traitement de documents. *Cahiers Leibniz*, (14), novembre 2000. 60, 63, 64, 65

[MM00] R. Mihalcea et D. Moldovan : Semantic indexing using wordnet senses. *In Proceedings of the ACL-2000 workshop on Recent advances in natural language processing and information retrieval : held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics - Volume 11*, RANLPIR '00, pages 35–45, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics. 65

[Moo48] C.N. Mooers : *Application of random codes to the gathering of statistical information.* Thesis, 1948. Thesis (M.S.) Massachusetts Institute of Technology. Dept. of Mathematics, 1948. 49

[MRS08] C.D. Manning, P. Raghavan et H. Schütze : *Introduction to information retrieval.* Cambridge University Press, Cambridge ; New York, 2008. 58, 61

[Nie08] J.-Y. Nie : Introduction à la RI : Indexation, 2008. 58

[Pin99] B. Pincemin : *Diffusion ciblée automatique d'informations : conception et mise en œuvre d'une linguistique textuelle pour la caractérisation des destinataires et des documents.* Thèse de doctorat, Paris IV, [S.l.], 1999. 56, 57, 58, 59, 62, 80, 81

[Por97] M. F. Porter : Readings in information retrieval. chapitre An algorithm for suffix stripping, pages 313–316. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997. 58

[PPL98] A. Pibarot, J. Picard et D. Labbé : Les syntagmes répétés dans l'analyse des commentaires libres. *S. Mellet (éd.) JADT*, pages 507–515, 1998. 60

[RJ08] S. Rosset et M. Jardino : Comparaison de documents : mesures de similarité et mesures de distance. june 2008. 68

[RK89] H. Ritter et T. Kohonen : Self-organizing semantic maps. *Biological Cybernetics*, 61:241–254, 1989. 10.1007/BF00203171. 64

[Sav93] J. Savoy : Stemming of french words based on grammatical categories. *Journal of the American Society for Information Science*, 44(1):1–9, 1993. 58

[Seb02] F. Sebastiani : Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47, 2002. 71

[SFW83] G. Salton, E.A. Fox et H. Wu : Extended boolean information retrieval. *Commun. ACM*, 26:1022–1036, November 1983. 52

[Sid02] S. Sidhom : *Plate-forme d'analyse morpho-syntaxique pour l'indexation automatique et la recherche d'information : de l'écrit vers la gestion des connaissances.* These, Université Claude Bernard - Lyon I, mars 2002. INSA DE LYON - EDIIS. 60

[SL83] A. Salem et P. Lafon : L'inventaire des segments répétés d'un texte. *Mots*, 6(1):161–177, 1983. 60

[SM83] G. Salton et M.J. McGill : *Introduction to modern information retrieval.* McGraw-Hill computer science series. McGraw-Hill, New York, 1983. 60, 61

[SPD+97] F. Sparacino, A. Pentland, G. Davenport, M. Hlavac et M. Obelnicki : City of news. *Ars Electronica Festival, Linz, Austria*, pages 8–13, 1997. 79

[ST09] G.M. Sacco et Y. Tzitzikas : *Dynamic Taxonomies and Faceted Search.* Springer, 2009. 49, 50

[TB89] P. Thoiron et H. Béjoint : Pour un index évolutif et cumulatif de cooccurrents en langue techno-scientifique sectorielle. *Meta*, 34(4): 661–671, 1989. 66

[Tom00] E.G. Toms : Understanding and facilitating the browsing of electronic text. *International Journal of Human-Computer Studies*, 52(3):423 – 452, 2000. 53, 54

[TR04] A. Tricot et J.-F. Rouet : Activités de navigation dans les systèmes d'information. *In Psychologie ergonomique : tendances actuelles*. Presses Universitaires de France - PUF, novembre 2004. 49

[Tri07] A.-P. Trinh : Classification de texte et estimation probabiliste par machine à vecteur support. *Actes de l'atelier DEFT07, CAp07*, 2007. 57

[VFL+06] A.M. Vercoustre, M. Fegas, Y. Lechevallier, T. Despeyroux *et al.* : Classification de documents xml à partir d'une représentation linéaire des arbres de ces documents. *et gestion des connaissances : EGC'2006*, 2006. 71

[Voo94] E.M. Voorhees : Query expansion using lexical-semantic relations. *In Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '94, pages 61–69, New York, NY, USA, 1994. Springer-Verlag New York, Inc. 65

Chapitre III

[BRN99] R. Baeza-Yates et B. Ribeiro-Neto : *Modern information retrieval*. 1999. 93

[BS08] M. Boughanem et J. Savoy : *Recherche d'information : état des lieux et perspectives*. Collection Recherche d'information et web, ISSN 1968-8008. Hermès science publ., Paris, 2008. 101

[CFLBC10] M.-L. Caron-Fasan, H. Lesca, A. Buitrago et A. Casagrande : Comment pérenniser un dispositif de veille anticipative à base de données numériques et textuelles : problématique et proposition. *In Actes colloque VSST'10*, 2010. 94

[CV04] Rudi Cilibrasi et Paul Vitanyi : Automatic meaning discovery using google. 2004. 96, 97

[CV06] R. Cilibrasi et P. Vitanyi : Automatic extraction of meaning from the web. pages 2309–2313, juillet 2006. 96

[CV07] R. Cilibrasi et P. Vitanyi : The google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, 19(3):370–383, 2007. 96

[LL11] H. Lesca et N. Lesca : *Les signaux faibles et la veille anticipative pour les décideurs : Méthodes et applications*. Hermes Science Publications, mai 2011. 94

[Mel01] S. Mellet : Lemmatisation et encodage grammatical : un luxe inutile ? *Lexicometrica*, (numéro spécial "Autour de la lemmatisation"), 2001. 93

[Tri07] A.-P. Trinh : Classification de texte et estimation probabiliste par machine à vecteur support. *Actes de l'atelier DEFT07, CAp07*, 2007. 93

Chapitre IV

[CFLBC10] M.-L. Caron-Fasan, H. Lesca, A. Buitrago et A. Casagrande : Comment pérenniser un dispositif de veille anticipative à base de données numériques et textuelles : problématique et proposition. *In Actes colloque VSST'10*, 2010. 124

[Pin02] B. Pincemin : Similarités texte texte. expérience d'une application de diffusion ciblée et propositions. *In Matematicas y Tratamiento de Corpus, Séminaire interlatin de linguistique appliquée*, page 35 52, 2002. 123

Chapitre V

[BMC11] M. Bodin, J.-C. Marty et T. Carron : Specifying collaborative tools in game-based learning environments : Clues from the trenches. *In Proc. of the 5th European Conference on Games Based Learning (ECGBL), Athens, Grece, 2011*, pages pp46–56, 2011. 172, 173

[HHD89] C. Hertzog, D.F. Hultsch et R.A. Dixon : Evidence for the convergent validity of two self-report metamemory questionnaires. *Developmental Psychology*, 25(5):687, 1989. 166

[MB04] A. Moschitti et R. Basili : Complex linguistic features for text classification : A comprehensive study. *Advances in Information Retrieval*, pages 181–196, 2004. 152

[Que47] R. Queneau : *Exercices de style*. Gallimard, 1947. 141, 159

[Seb02] F. Sebastiani : Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47, 2002. 152

[Val12] F. Vallet : *Comment évalue-t-on l'efficience de notre mémoire : Le rôle des attributions causales et des théories naïves*. Thèse de doctorat, Université de Grneoble, 2012. 166

[Yan99] Y. Yang : An evaluation of statistical approaches to text categorization. *Information retrieval*, 1(1):69–90, 1999. 152

Annexes

1 Annexes du chapitre I

1.1 FULL texts sur le domaine de l'agroalimentaire utilisés dans la méthode Puzzle

Deux maladies menacent les récoltes de cacao en Afrique et au Brésil

Afrique et Brésil - Des maladies, encore méconnues des scientifiques, pourraient causer la perte d'un tiers des récoltes dans ces pays où la production de cacao est la plus importante.

Ces dernières années, les cacaotiers ont été plantés très près les uns des autres, à cause de la demande croissante à laquelle les producteurs doivent faire face. De fait, la diversité des autres espèces d'arbres qui poussent habituellement entre eux a fortement diminué. Ce phénomène, couplé à la plantation d'arbres en milieu sec, a accru le risque de contamination des cultures.

Deux causes ont été identifiées comme étant à l'origine de la perte des récoltes. En Afrique occidentale, le Cocoa Swollen Shoot Virus, ou CSSV, tue les arbres. Au Brésil, c'est un champignon appelé « balai des sorcières » qui menace les récoltes.

Le CSSV est facilement transporté par les insectes et il est donc très difficile à éradiquer. La seule solution trouvée pour éviter qu'il ne s'étende aux cultures voisines consiste à détruire les arbres touchés et à protéger ceux qui sont sains.

Une espèce de cacaotiers partiellement résistante au CSSV a déjà été découverte en Afrique et les scientifiques tentent de la rendre plus forte. De son côté, le service américain de l'agriculture de Miami, en Floride, tente de mettre le doigt sur des gènes résistants, présents chez d'autres espèces d'arbres.

13/04/2009
http ://www.maxisciences.com/agriculture/deux-maladies-menacent-les-recoltes-de-cacao-en-afrique-et-au-bresil_art1519.html

Monsanto haalt bakzeil bij Duitse kortgedingrechter

Op 14 april heeft de Duitse minister van Landbouw Ilse Aigner een vrijwaringsclausule geactiveerd tegen de teelt van MON 810. Een Duits rechtscollege heeft nu een beroep in kort geding vanwege de Amerikaanse agroreus Monsanto tegen de door Berlijn verboden teelt van de genetisch gewijzigde maïs verworpen.

Aigner baseerde zich bij het verbod op twee nieuwe studies die "nieuwe wetenschappelijke elementen" aan het licht brachten. Met name dat het door Monsanto in het zaad ingevoerde gen schadelijk zou zijn voor lieveheersbeestjes en vlinders. Monsanto is tegen de beslissing in beroep gegaan.

Maar het beroep in kort geding werd door de administratieve rechtbank van Braunschweig verworpen omdat na een "risicosituatie een dergelijk verbod rechtvaardigt, zoals de wet op de biotechnologie voorziet". Opdat zo'n beslissing legitiem zou zijn, is het niet nodig dat er een duidelijk identificeerbaar gevaar bestaat, het volstaat dat er aanwijzingen in die zin zijn, zo meent het rechtscollege.

Monsanto kan hiertegen in beroep gaan en liet weten "de mogelijkheid te bestuderen om nieuwe argumenten voor te leggen". In elk geval komt er in Braunschweig nog een procedure ten gronde, waarbij er mondelinge debatten zullen zijn. Voor dit geding is nog geen datum vastgelegd.

Met het verbod schaarde Duitsland zich aan de zijde van Frankrijk, Griekenland, Oostenrijk, Hongarije en Luxemburg. Die landen hebben de teelt van MON810 eerder reeds gebannen, waarbij ze zich beroepen op het voorzorgsprincipe.

06/05/2009
http ://www.vilt.be/Monsanto_haalt_bakzeil_bij_Duitse_kortgedingrechter

Wetsontwerp legt bijmengplicht biobrandstoffen vast

De ministerraad heeft een wetsontwerp goedgekeurd dat producenten van benzine en diesel verplicht om in hun brandstoffen 4 procent biodiesel of bio-ethanol te mengen. Ons land moet zo in de buurt komen van de Europese streefcijfers.

Europa wil dat motorbrandstoffen vanaf volgend jaar voor minstens 5,75 procent uit biobrandstoffen bestaan. Zonder bijkomende maatregelen benadert ons land die streefcijfers zelfs niet. De regering opteerde er in 2006 voor om biobrandstoffen te stimuleren via fiscale vrijstellingen, maar dat leverde zo goed als niets op. Vorig jaar werd slechts voor 1 procent gemengd. Vrijwel nergens in België kan biobrandstof worden getankt.

De situatie zorgde ervoor dat een aantal bedrijven die de voorbije jaren zwaar investeerden in biobrandstoffen hun producten niet verkocht kregen en daardoor zwaar in de problemen zitten. Minister van Energie Paul Magnette werkte daarom een regeling uit.

Het ontwerp, dat nu naar de Raad van State wordt gestuurd voor advies, verplicht de oliemaatschappen om 4 procent biobrandstoffen te mengen in diesel en benzine.

Eerder liet Magnette weten dat de verplichting voorlopig voor twee jaar geldt en indien nodig kan worden verlengd.

De fiscale vrijstelling van accijnzen blijft bestaan. De producenten kunnen er wel alleen maar van genieten wanneer ze hun biobrandstoffen kopen bij bedrijven die in België gevestigd zijn.

09/05/09
http ://www.vilt.be/Wetsontwerp_legt_bijmengplicht_biobrandstoffen_vast

Mexicaanse griep treft export Amerikaans varkensvlees

Vleesproducenten in de VS lijden ernstig schade door de uitbraak van Mexi-caanse griep. Tenminste tien landen hebben de import van Ameri-kaanse varkens en varkensvlees verboden. In de meeste gevallen gaat het om een importstop van producten uit zowel Mexico als de Amerikaanse staten waar Mexicaanse griep is vastgesteld.

Onder deze landen bevinden zich grote importeurs als Rus-land en China. Rus-land besloot zelfs de import van niet-verhit pluimvee en pluimveevlees aan banden te leggen. Het H1N1-virus bevat immers niet alleen muta-ties van varkensgriep, maar ook menselijke en vogelgriep. Rusland en China beroepen zich op de BSE–crisis in 2000. Toen is aanvanke-lijk gesteld dat vlees uit getroffen gebieden veilig was, wat later niet zo bleek te zijn. De beide landen zijn samen goed voor bijna een derde van de Ameri-kaanse uitvoer aan varkensvlees.

De schade loopt op tot circa 203 miljoen euro, zo ver-klaarde een onderzoeker van de University of Missouri op CNN. Het ministerie van Landbouw acht het mogelijk dat een kwart tot een derde van de varkenshou-ders in de problemen kan komen. De economische schade is nog groter wanneer wordt gekeken naar de gevolgen voor de land-bouw in bredere zin. De varkens-vleessector verbruikt volgens on-derzoekers van Purdue Universi-ty 28 procent van de granen die gevoerd wordt in de veehouderij. De prijzen hiervan zijn deels gekelderd door een verwachte afna-me van de vraag.

In de VS is vooral vleesprodu-cent Smithfield Foods getroffen. Een groot var-kensbedrijf in La Gloria is de verdachte haard van het virus, en half eigendom van Smithfield Foods. Deze firma is wereldmarktleider met een om-zet van 9 miljard euro. Keiji Fukuda van de Wereldgezondheidsorganisatie (WHO) heeft bij herha-ling verklaard dat het virus niet kan worden over-gedragen door vlees. Volgens de VS is bij nog geen enkel varken Mexicaanse varkensgriep vastge-steld. Het is dan ook waarschijn-lijk dat de VS en Mexico een klacht indienen bij de Wereldhan-delsorganisatie WTO.

02/05/09
http ://www.vilt.be/Mexicaanse_griep_treft_export_Amerikaans_varkensvlees

« Appelvoorraden in koelruimtes nog enorm »

De export naar Rusland is sterk bepalend voor de appel- en perenprijzen. Net zoals de boeren voelt ook fruitexporteur Kris Wouters de devaluatie van de roebel aan den lijve. "Onze verkoop aan Rusland is normaliter goed voor vijftig vrachtwagens van twintig ton per week. Maar op dit ogenblik laden we nog de helft van de volumes van vorig jaar", luidt het.

"De roebel is tot veertig procent gedevalueerd, waardoor ons fruit veel duurder geworden is voor de Russen. Bovendien kende Polen een overvloedige oogst. Omdat ook de Poolse munt gedevalueerd is, kunnen zij veel goedkoper leveren dan wij. De transportkosten liggen ook veel lager", weet Kris Wouters.

Zijn fruithandel is al sinds 1993 actief in Rusland. Sindsdien heeft de exporteur uit Rummen er een ruim cliënteel opgebouwd. Naast appelen en peren transporteert hij ook aardbeien en tomaten. Naast de productie van zijn eigen plantages verkoopt Wouters ook de oogst van producenten uit Haspengouw en het Hageland door aan Rusland. Zelfs Fransen en Spanjaarden doen beroep op zijn exportbedrijf.

Maar door de lage prijzen weten de boeren momenteel niet goed wat te doen. "Telers wachten af om hun frigo's leeg te maken", vertelt Wouters. "Ze hopen dat de stocks in Polen binnenkort op zijn. Dan kunnen ze de Russische markt weer innemen en zullen de prijzen wat stijgen. Het blijft bang afwachten, ook voor volgend seizoen".

Eén van de bedrijven die nog over een grote voorraad appelen beschikken, is het fruitteeltbedrijf Vanhellemont in Meensel-Kiezegem. In de koelcellen zit nog ongeveer 250 ton, oftewel tien volle vrachtwagens. "De helft van mijn oogst zit nog in de koelcel", zucht Mario Vanhellemont. "De vraag naar onze appelen is wel groot, maar de prijs die ervoor betaald wordt, is veel te laag".

De fruitboer vreest dat de prijzen de komende weken niet meer zullen klimmen. "Eigenlijk hebben we nu geen keuze meer. De appelen moéten weg, want in augustus beginnen we met de nieuwe oogst. We verkopen dus, maar wel fors onder onze productiekost.

04/05/09
http ://www.vilt.be/Appelvoorraden_in_koelruimtes_nog_enorm

AB InBev fait monter la pression sur ses fournisseurs

AB InBev a décidé unilatéralement, dès janvier, de n'honorer ses factures qu'après un délai de 120 jours, contre 30 jours auparavant. Les fournisseurs sont mécontents mais affirment qu'ils ne peuvent rien faire sous peine de perdre leur contrat, rapporte De Tijd dans son édition de samedi.

AB InBev a, dans un bref communiqué, confirmé son intention de porter ses délais de paiement à 120 jours à partir de la réception de la facture : «La situation économique actuelle a poussé AB Inbev - tout comme beaucoup d'autres multinationales - à revoir ses délais et ses conditions de paiement», a indiqué Karen Couck,

responsable de la communication du groupe.

Concrètement, AB Inbev a décidé de porter ses délais de paiement à 120 jours. En Belgique, ce délai fait partie des négociations commerciales entre un fournisseur et son client. Le groupe brassicole rappelle dans la foulée qu'il respecte l'ensemble des lois et des règles, y compris ses obligations contractuelles.

AB InBev tente par tous les moyens de réduire son endettement de 45 milliards de dollars, contracté lors du rachat d'Anheuser-Busch par InBev l'an dernier. Un montant de 7 milliards de dollars doit être remboursé pour novembre 2009. La moitié l'a déjà été grâce à l'émission d'obligations en janvier, et le groupe a empoché 667 millions de dollars de la vente de ses parts dans la brasserie chinoise Tsingtao. Outre des cessions d'actifs, l'idée est de réduire de 2,25 milliards de dollars ses dépenses annuelles.

De son côté, Vincent Van Quickenborne, ministre pour l'Entreprise, a décidé de demander une enquête informelle au Conseil de la concurrence sur un éventuel abus de position dominante. Car ces nouvelles conditions sont apparemment «à prendre ou à laisser» et les fournisseurs, souvent des entreprises beaucoup plus petites que le groupe brassicole, sont obligées de suivre.

Le brasseur louvaniste utiliserait donc sa taille - il contrôle un quart du marché global de la bière - «pour changer les règles du jeu (Ndlr, avec ses fournisseurs), écrivait ainsi Gerard Rijk, analyste chez ING à Amsterdam, dans une note datée du 30 mars (citée par Bloomberg). Comme l'industrie de l'orge et du malt est encore très fragmentée, le pouvoir de négociation des fournisseurs face à AB InBev est limité.»

«Il s'agit d'une enquête préliminaire informelle, destinée à vérifier qu'il y a matière à ouvrir une enquête formelle, prévient Tim Van Broeckhoven, porte-parole du ministre, cité par Bloomberg. Selon lui, le Conseil de la concurrence pourrait discuter du dossier avec la Commission européenne, en fonction des résultats de son enquête préliminaire.

20/04/09
http ://trends.levif.be/economie/actualite/entreprises/ab-inbev-fait-monter-la-pression-sur-ses-fournisseurs/article-1194637279504.htm

Bientôt l'étiquette "Nourri sans OGM" ?

Une proposition de loi écolo vise à mieux informer le consommateur.

En Belgique comme dans toute l'Union européenne, un aliment qui contient plus de 0,9 % d'OGM doit être dûment étiqueté comme tel. Ces produits (quelques huiles de soja ou de maïs) sont toutefois minoritaires dans nos supermarchés. La population rechigne à mettre du transgénique dans son assiette.

Mais qu'en est-il pour la viande, le lait, les œufs et autres produits issus d'animaux qui consomment, eux, des OGM ? En fait, rien n'est prévu. Et pourtant, selon

Thérèse Snoy, députée fédérale Ecolo, environ 40 % de la nourriture des animaux d'élevage en Belgique est transgénique. Elle va donc déposer à la Chambre une proposition de loi qui vise à permettre aux producteurs d'étiqueter leurs aliments "Sans OGM", ce qui offrirait une garantie au consommateur sur ce qu'il achète. Une mesure qui, si elle venait à être appliquée, serait loin d'être anodine dans la mesure où elle nécessiterait la mise en œuvre d'un système de traçabilité des OGM dans la nourriture animale.

06/05/09
http ://www.lalibre.be/actu/belgique/article/500272/bientot-l-etiquette-nourri-sans-ogm.html

Polémique autour des peupliers OGM (suite de l'article précédent)

Par ailleurs, ce mercredi, les députés Ecolo déposeront une autre proposition de loi dont l'objectif est de donner aux ministres plus de latitude face aux avis scientifiques lorsqu'il s'agit d'autoriser ou non une culture d'OGM en plein air.

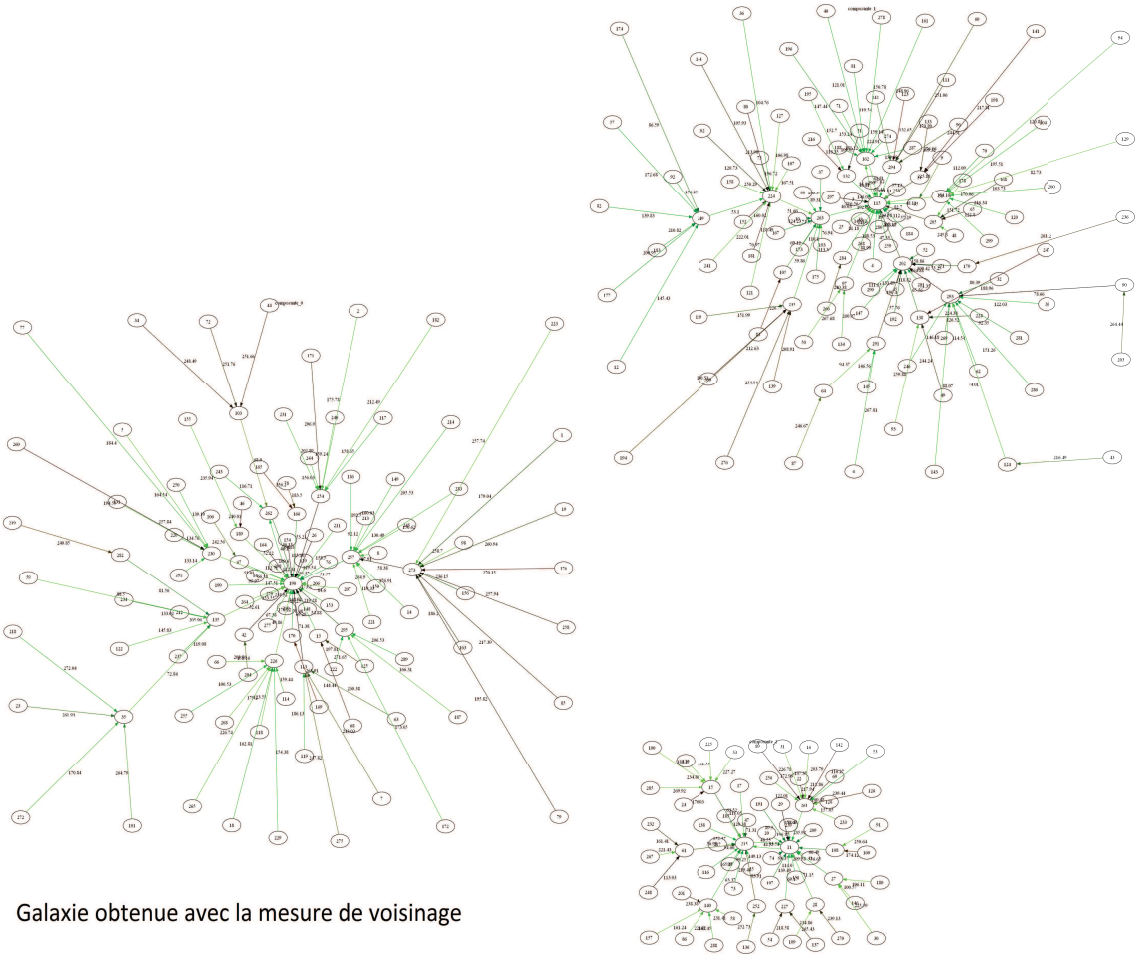
Ce n'est pas tout à fait un hasard, c'est aussi ce mercredi que Patricia Ceysens, la ministre flamande en charge notamment des Sciences et de l'Innovation, doit planter les fameux peupliers du Vlaams Instituut voor Biotechnologie.

Ces arbres, dont le patrimoine génétique a été modifié pour faciliter la production de biocarburant, ont en effet été, ces derniers mois, au cœur d'une polémique, la demande (flamande) ayant été dans un premier temps refusée par les ministres (francophones) de la Santé et de l'Environnement. Ministres dont la décision a ensuite été suspendue par le Conseil d'Etat, parce que, pour faire court, ils s'étaient opposés sans contre-arguments scientifiques à l'avis du conseil consultatif de biosécurité. Pour le Conseil d'Etat, en la matière, les arguments économiques ou éthiques avancés ne tenaient pas. La proposition Ecolo vise dont à inclure au sein du conseil de biosécurité une chambre spéciale qui évaluerait les demandes d'autorisation sous un prisme "socio-éthico-économique" et non plus seulement scientifique.

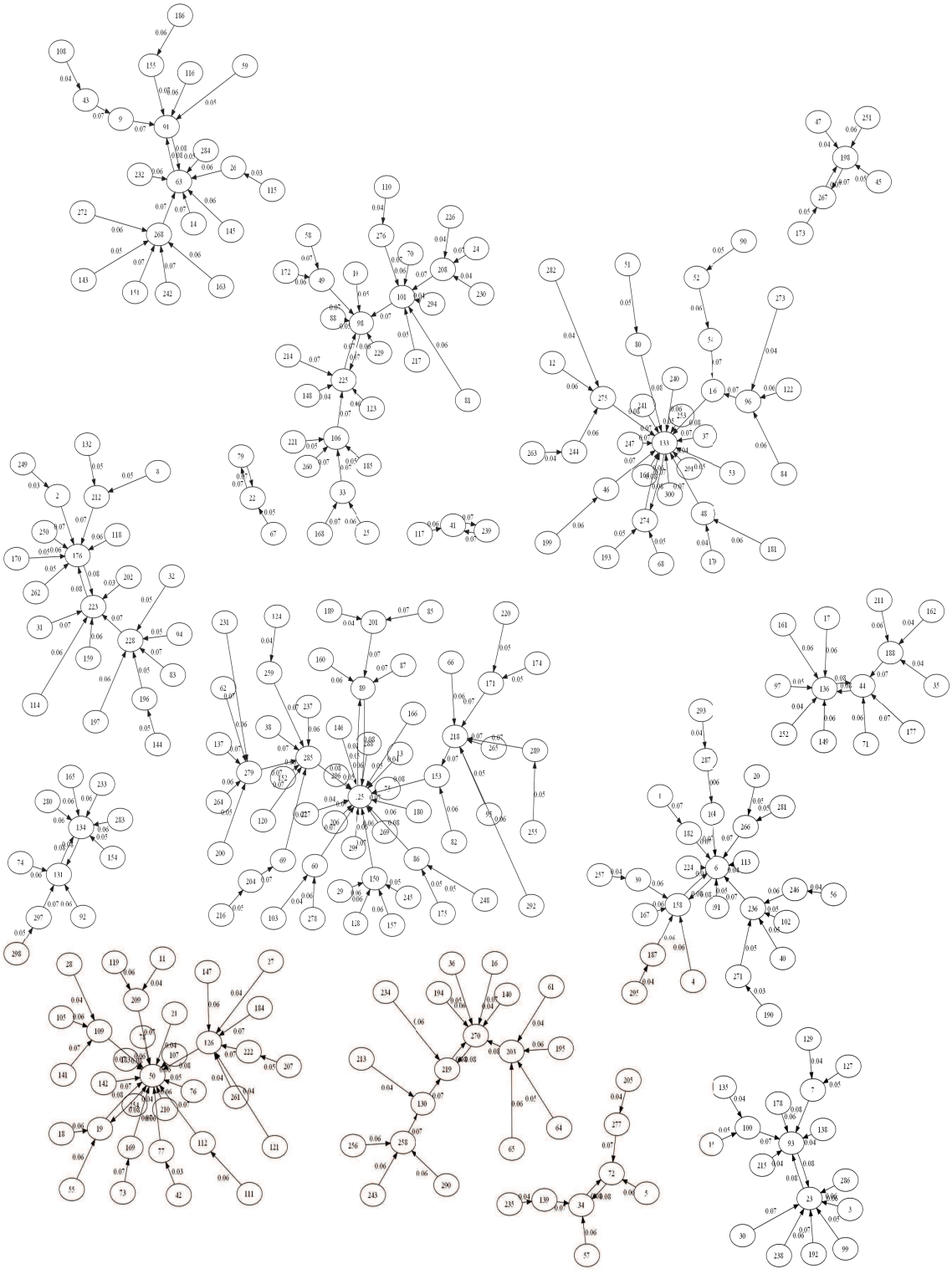
06/05/09
http ://www.lalibre.be/actu/belgique/article/500272/bientot-l-etiquette-nourri-sans-ogm.html

2 Annexes du chapitre III

2.1 Galaxies liées aux textes aléatoires



Galaxie obtenue avec la mesure de voisinage



Galaxie obtenue avec cosinus sur des vecteurs booléens

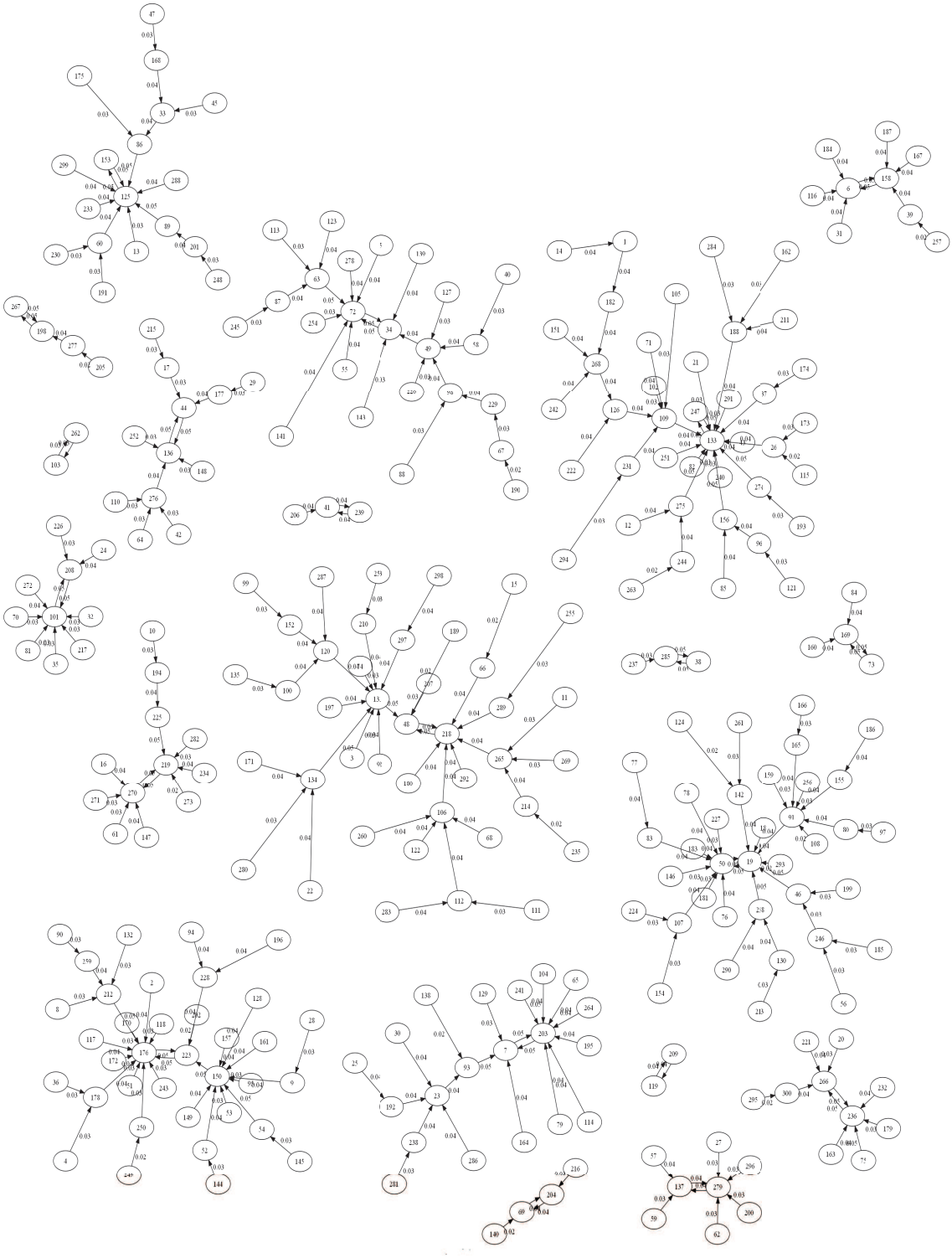
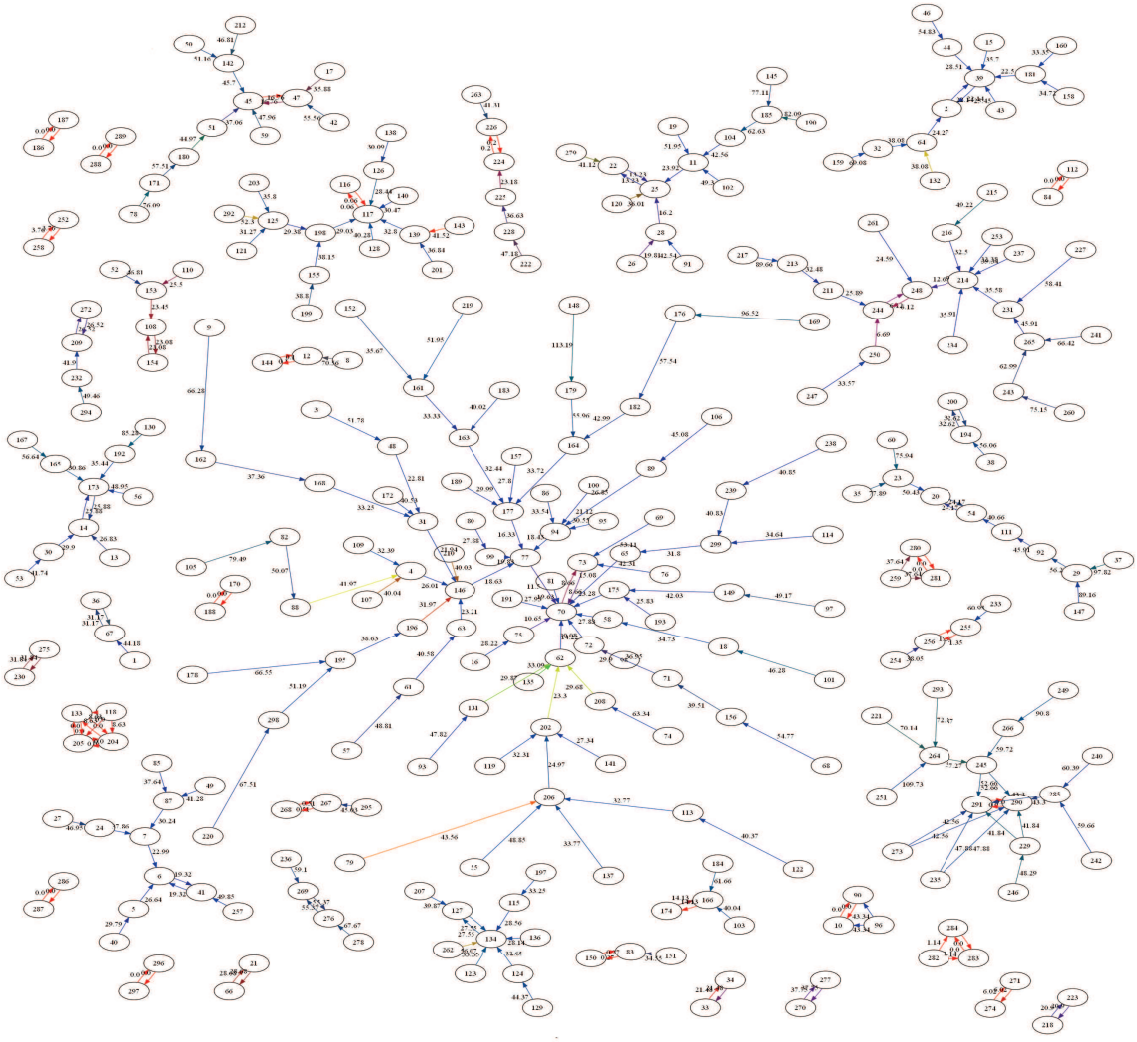
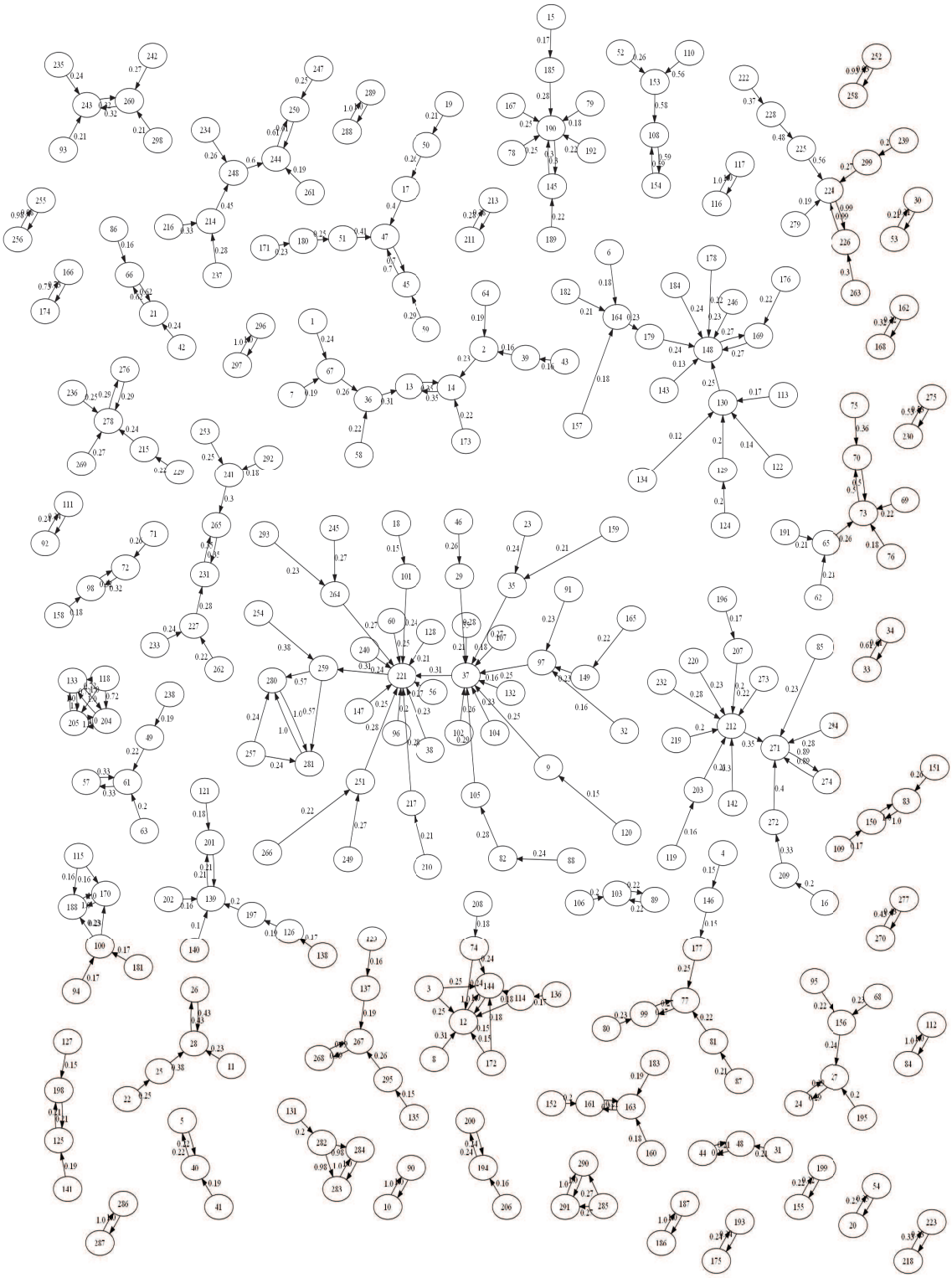


Figure A.1 – Textes aléatoires

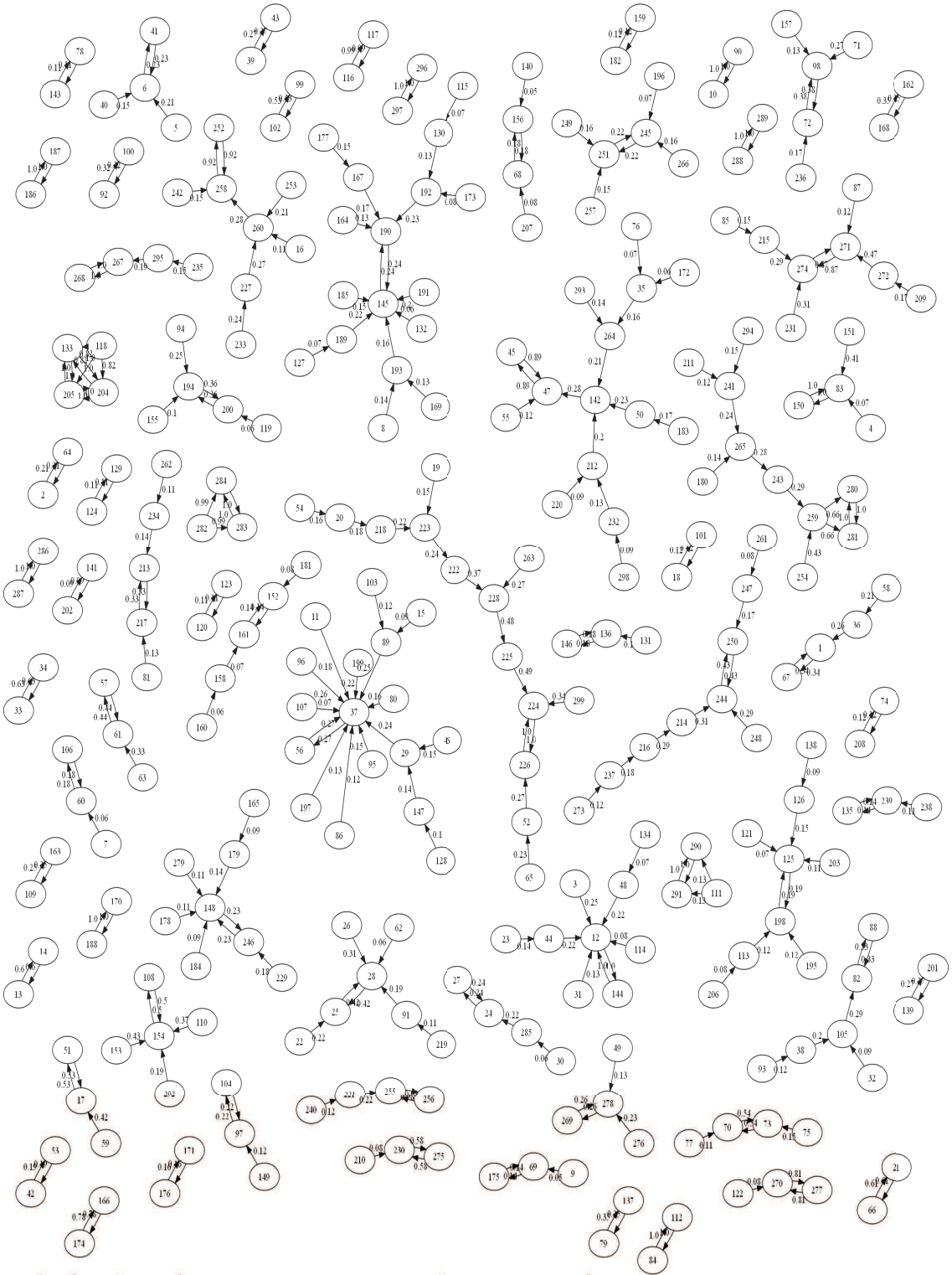
2.2 Galaxies liées aux textes sur le CO_2



Galaxie obtenue avec la mesure de voisinage

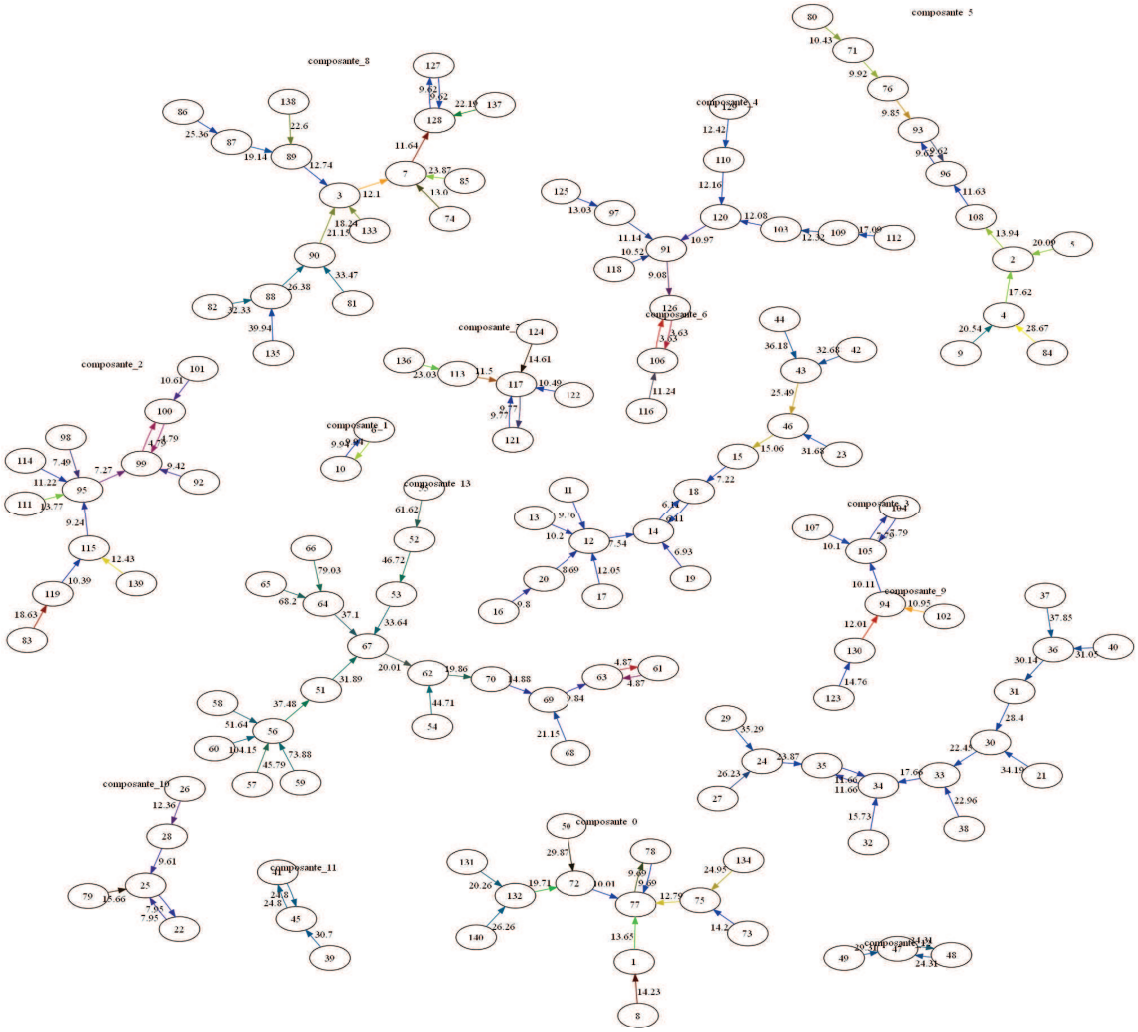


Galaxie obtenue avec cosinus sur des vecteurs booléens

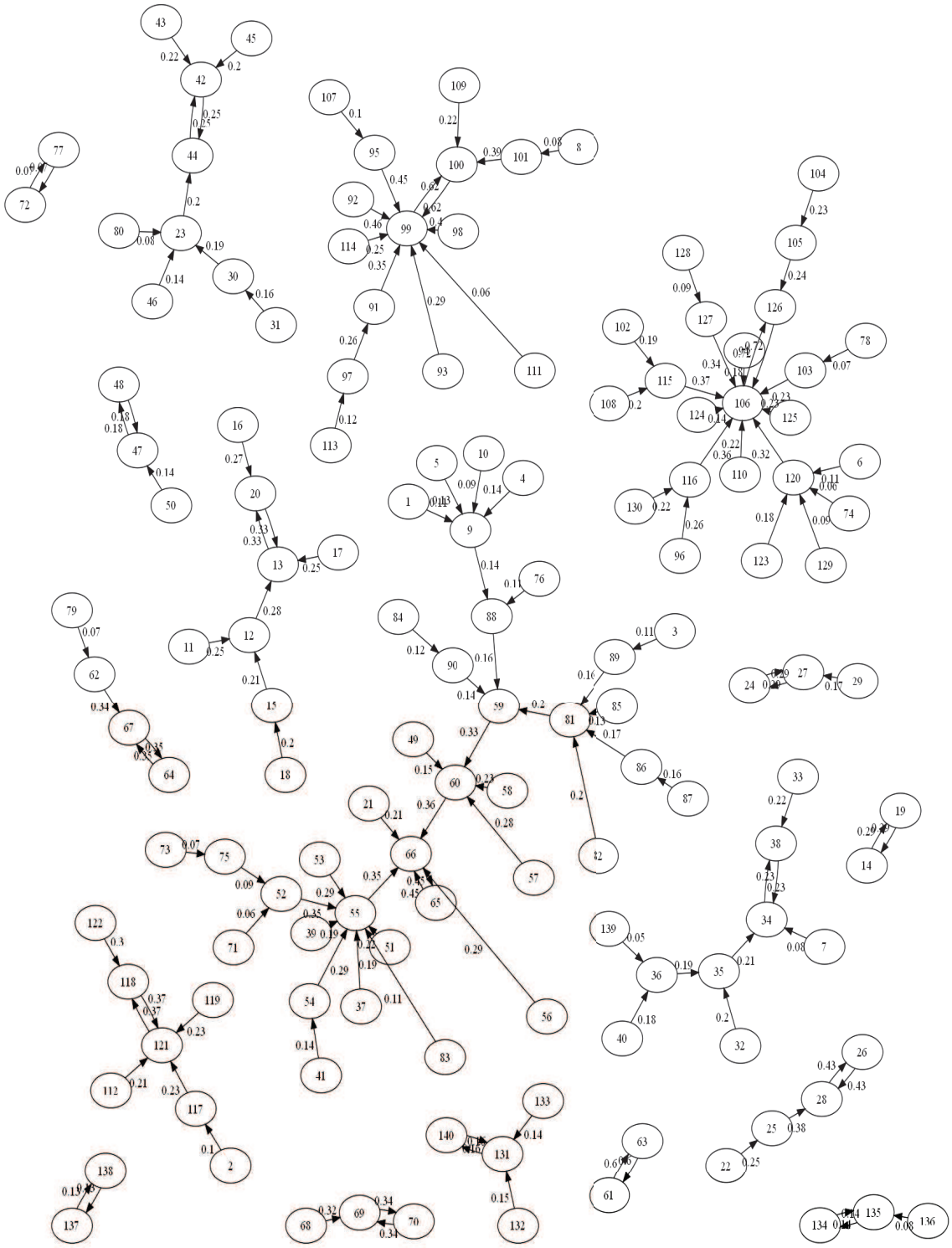


Galaxie obtenue avec cosinus sur des vecteurs TF-IDF

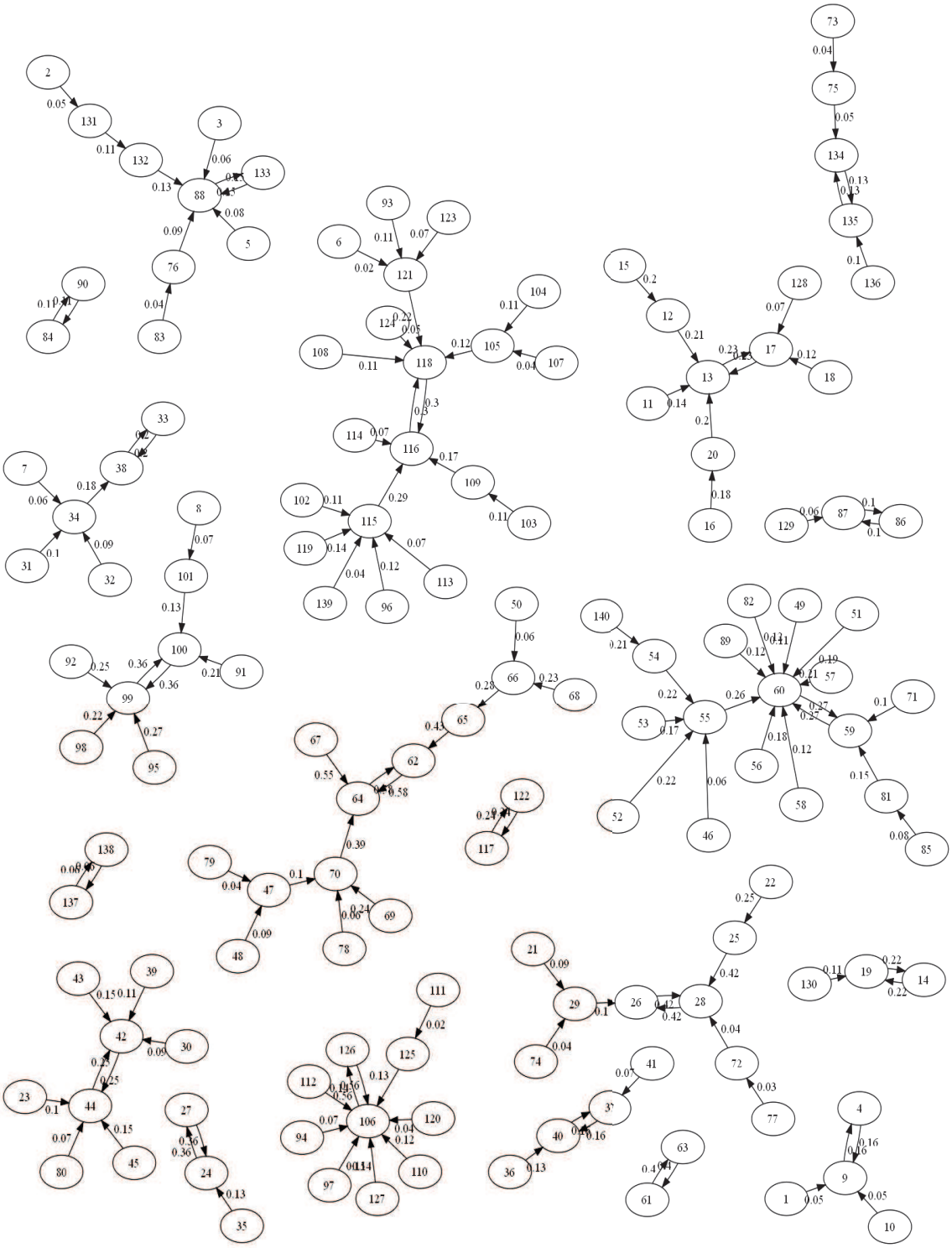
2.3 Galaxies liées aux textes variés



Galaxie obtenue avec la mesure de voisinage



Galaxie obtenue avec cosinus sur des vecteurs booléens



Galaxie obtenue avec cosinus sur des vecteurs TF-IDF

3 Annexe du chapitre V

3.1 Test pour la comparaison de deux proportions

Le descriptif de ce test est tiré de : F. Santos, *Tests pour la comparaison de deux proportions*, février 2011, http ://frederic.santos.perso.sfr.fr/pdf/stat/test_prop.pdf.

On suppose qu'on a des données de comptage d'une même caractéristique dans deux populations différentes — par exemple, le nombre d'individus ayant une certaine maladie. On cherche à savoir si les proportions d'individus présentant cette caractéristique peuvent être considérées comme identiques entre ces deux populations.
On suppose que la population 1 compte n_1 individus, et que la population 2 compte n_2 individus. Les données recueillies peuvent s'interpréter comme des réalisations de deux familles de variables aléatoires binaires :

1. On note $X_1, ..., X_{n_1}$ les résultats obtenus dans la première population. Pour chaque individu i de cette population, X_i vaut 1 s'il est malade, et vaut 0 sinon. La probabilité théorique d'être malade quand on est membre de cette population est notée p_1, de telle sorte que chaque variable aléatoire X_i suit une loi de Bernoulli $\mathcal{B}(p_1)$.
2. De même, on note $Y_1, ..., Y_{n_2}$ les résultats obtenus dans la deuxième population. On note p_2 la probabilité théorique d'être malade dans cette population. Chaque Y_i suit alors une loi de Bernoulli $\mathcal{B}(p_2)$.

Si on a compté X individus malades dans la population 1 et Y individus malades dans la population 2, alors par définition :

$$X = \sum_{i=1}^{n_1} X_i \sim \mathcal{B}(n_1; p_1) \tag{1}$$

$$Y = \sum_{i=1}^{n_2} Y_i \sim \mathcal{B}(n_2; p_2) \tag{2}$$

Si les effectifs n_1 et n_2 sont suffisamment grands, on peut faire l'approximation normale suivante :

$$\frac{X}{n_1} \sim \mathcal{N}(p_1; \frac{p_1(1-p_1)}{n_1}) \tag{3}$$

$$\frac{Y}{n_2} \sim \mathcal{N}(p_2; \frac{p_2(1-p_2)}{n_2}) \tag{4}$$

On en déduit en particulier la loi de la différence entre les proportions empiriques

observées au sein des deux populations :

$$\frac{X}{n_1} - \frac{Y}{n_2} \sim \mathcal{N}(p_1 - p_2; \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}) \tag{5}$$

On cherche maintenant à savoir si l'hypothèse d'égalité des proportions théoriques p_1 et p_2 est raisonnable au vu de ce qui est observé — autrement dit, s'il est raisonnable d'affirmer qu'aucune des deux populations n'est plus amenée à être malade que l'autre. On effectue donc le test d'hypothèses suivant :

$$\mathcal{H}_0 : p_1 = p_2 \tag{6}$$

$$\mathcal{H}_1 : p_1 \neq p_2 \tag{7}$$

Sous $\mathcal{H}_0$, on a $p_1 = p_2 = p$. Alors :

$$\frac{X}{n_1} - \frac{Y}{n_2} \sim \mathcal{N}(0; p(1-p)(\frac{1)}{n_1} + \frac{1)}{n_2})) \tag{8}$$

On estime p par la proportion totale observée de malades sans distinction de populations :

$$\hat{p} = \frac{X+Y}{n_1+n_2} \tag{9}$$

On en déduit finalement le test paramétrique approché de niveau α : on rejette $\mathcal{H}_0$ si

$$\frac{X}{n_1} - \frac{Y}{n_2} > u_{1-\frac{\alpha}{2}} \sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})} \tag{10}$$

où $u_{1-\frac{\alpha}{2}}$ est le quantile bilatéral d'ordre α de la loi $\mathcal{N}(0;1)$.

La veille anticipative stratégique et intelligence collective (VASIC) proposée par Lesca est une méthode aidant les entreprises à se mettre à l'écoute de leur environnement pour anticiper des opportunités ou des risques. Cette méthode nécessite la collecte d'informations. Or, avec le développement des technologies de l'information, les salariés font face à une surabondance d'informations. Afin d'aider à pérenniser le dispositif de veille stratégique, il est nécessaire de mettre en place des outils pour gérer la surinformation. Dans cette thèse, nous proposons une mesure de voisinage pour estimer si deux informations sont proches ; nous avons créé un prototype, nommé Alhena, basé sur cette mesure. Nous démontrons les propriétés de notre mesure ainsi que sa pertinence dans le cadre de la veille stratégique. Nous montrons également que le prototype peut servir dans d'autres domaines tels que la littérature, l'informatique et la psychologie.
Ce travail est pluridisciplinaire : il aborde des aspects de veille stratégique (en sciences de gestion), de la recherche d'informations, d'informatique linguistique et de mathématiques. Nous nous sommes attachés à partir d'un problème concret en sciences de gestion à proposer un outil qui opérationnalise des techniques informatiques et mathématiques en vue d'une aide à la décision (gain de temps, aide à la lecture,...).

Mots clefs : veille stratégique anticipative, mesure de voisinage, intelligence collective, similarité, recherche d'information

Business environmental scanning and collective intelligence (VASIC) as proposed by Lesca is a method to help companies tune in to their environment to anticipate opportunities or risks. This method requires collecting information, yet with the development of information technology, employees face a glut of information. To help sustain VASIC, it is necessary to develop tools to manage information overload. In this thesis, we propose a nearness measurement to estimate if two pieces of information are similar and we have created a prototype, called Alhena, based on this measurement. We demonstrate the properties of our measurement and its relevance in the context of VASIC. We also show that the prototype can be used in other fields such as literature, computer science and psychology.
This work is multidisciplinary as it covers aspects of business environmental scanning (management science), research information, computer linguistics and mathematics. We focus on a concrete problem in management science to provide a tool that operationalizes computational and mathematical techniques with a goal of providing decision making support (time saving, reading assistance, ...).

Keywords : Business environmental scanning, nearness measurement, , similarity, information retrieval