BELUGA : un outil pour l'analyse dynamique des connaissances de la littérature scientifique d'un domaine. Première application au cas des maladies à prions.

Nicolas Turenne*, Marc Barbier**

*Biométrie et Intelligence Artificielle (BIA), INRA
78 352 Jouy-en-Josas
turenne@jouy.inra.fr
**Systèmes Agraires et Développement (SAD), INRA
78 850 Thiverval-Grignon
barbier@grignon.inra.fr

Résumé. Un projet ciblé sur l'étude du domaine des maladies à prions a permis de formaliser une méthodologie commune, sociologique et informatique, de compréhension de sa dynamique par l'analyse thématique. Nous avons créé une plate forme d'indexation de notices bibliographiques dont le but est d'extraire des associations évoluant à travers des intervalles de temps. Beluga propose une chaîne de traitement basée sur l'indexation des documents en unités de base: références, auteurs, termes simples et composés, organismes. L'outil est fondé sur une double approche d'apprentissage et de visualisation qui automatise les processus d'extraction de groupes d'auteurs et de termes, et permet à l'utilisateur de revenir aux données documentaires sources. L'analyse diachronique de corpus de documents électroniques nous permet d'analyser comment la terminologie est structurée en thématiques en nous ramenant au problème de détection des tendances thématiques émergentes.
Mots-clés: fouille de textes; extraction terminologique, traitement de corpus; extraction de connaissances à partir de corpus; détection de thématiques émergentes; maladies à prions ; scientométrie.

1. Introduction

L'utilisation par les chercheurs en sciences sociales d'outil de fouille, d'extraction de connaissances et d'aide à l'interprétation est d'actualité pour caractériser les dynamiques d'un certain nombre « d'affaires », dont le dossier maladies à prions [Chateauraynaud et Torny, 1999]. Notre travail s'inscrit dans cette perspective[1]. Dans le dossier des maladies à prions, la complexité des dynamiques sociales est forte puisque de nombreux mondes professionnels sont impliqués et en interactions. L'une des « couches » d'information importante est bien sûr celle constituée par les bases de données scientifiques.

Concernant les maladies à prions, les travaux scientifiques sur la dynamique de ce domaine sont, en France, finalement peu nombreux au regard de l'importance des crises sanitaires liées à la vache-folle. [Barbier et De Looze, 1999], [Maunoury et al., 1999]. Le

1 *L'expertise du Comité interministériel sur les ESST à la frontière entre recherche et décision publique* (resp. scientifique: Marc Barbier, INRA), GIS Prions *AO 2001.*

but du travail que nous présentons ici est de chercher à automatiser partiellement l'acquisition de connaissances par une analyse distributionnelle, en testant : l'applicabilité des méthodes à des volumes importants de données, le couplage de méthodes d'apprentissage statistique et de traitements automatiques de la langue, et l'utilisation de méthodes dans un cadre orienté utilisateur à des fins d'utilité des résultats et d'évaluation guidée par l'expertise. Nous présentons ainsi l'outil *Beluga* développé pour décrire et analyser la dynamique de constitution d'un domaine de recherche.

2. L'analyse thématique : état de l'art

L'identification automatique de thèmes dans les documents électroniques est un sujet ancien [Stone et Kelly, 1966] mais les méthodes ont été rarement appliquées à l'étude de crises sanitaires. Certaines approches font état d'études bibliométriques sur les *Newsgroups* [Bar-Ilan, 1997] ou se penchent sur une méthode de visualisation de clusters [Chen et a l. 2002]. Mais aucune discussion n'est finalement développée sur la description des thèmes scientifiques au cours du temps.

De façon plus générale, rares sont les approches qui prennent en compte le temps de manière explicite. On trouve traditionnellement des méthodes d'analyse thématique qui associent un mot ou un segment à une catégorie, ou des méthodes qui comparent l'état d'un descripteur avec son voisinage par rapport à ce même voisinage à une époque précédente [Lelu et Ferhan, 1998]. L'approche décrite ici s'inspire des méthodologies de l'accès à l'information textuelle afin d'intégrer la notion d'indice de dynamique temporelle des connaissances. Dans des espaces de variables multidimensionnelles, l'extraction des connaissances est plus simple en considérant une représentation condensée dispersée en réseau. Nous avons choisi la technique de classification non supervisée dans cet objectif. Deux techniques sont mises alors en confrontation, l'une l'utilisant le contexte adjacent [Agrawal et Srikant, 1994] et l'autre le contexte non adjacent [White & Griffith, 1981].

L'analyse de schémas dans les motifs est étudiée depuis peu [Li et al., 2002]. Ces travaux étudient l'aspect périodique ou calendaire d'itemsets ou de règles qui apparaîtraient au cours du temps. Dans une perspective exploratoire nous souhaitons produire des motifs qui, non seulement pourront être visualisés explicitement par l'utilisateur avec leur évolution au cours du temps, mais aussi une prise en compte, en tant qu'unité, pour l'estimation d'un indice.

La méthode de co-citations d'auteurs que nous utilisons produit une suite de classes au même titre que les motifs basés sur une technique de treillis de concepts [Ganter et Wille, 1999]. L'obtention de classes sur différents intervalles permet d'observer l'apparition ou la disparition de « groupes » de descripteurs pour ensuite croiser ces groupes avec les termes du domaine et ainsi faire apparaître les thèmes associés.

3. Données

Deux types de données sont disponibles: les données brutes et des connaissances externes. Les données brutes constituent un entrepôt de données textuelles à partir duquel on cherchera à extraire les tendances thématiques. Les connaissances externes sont des hiérarchies de connaissances du domaine biomédical et de l'ESB qui peuvent servir à guider ou amorcer une analyse thématique.

Nous rappelons qu'un corpus est une collection de documents qui décrit le plus largement possible un domaine de connaissances. Cette définition reste vague: elle ne renseigne pas sur la nature du document, sur la couverture du domaine et sur la taille, mais elle fixe le type de base de données que nous allons traité. Les corpus forment un entrepôt d'informations à partir duquel notre objectif est de dégager des connaissances sur le domaine qui sont, au départ, descriptives et textuelles. Un corpus de notices apporte ses contraintes structurelles propres. La langue est fixée: l'anglais. Chaque notice est composée de champs attributs-valeurs: *auteur*, *organisation*, *date*, *résumé*, *mot-clés* (Remarques: certains champs, comme *résumé* ou *mots-clés*, sont absents de certaines notices, le champ *date* n'est composé que de l'année). Le corpus ESB a été formé principalement de notices scientifiques MEDLINE et du SCI récupérées à partir d'une requête élaborée avec les noms de pathologie en anglais ou d'agents des ESST (par exemple: *BSE, bovine spongiform encephalopathy, CJD, Creutzfeldt-Jakob, prion, scrapie*).

4. Méthodes

4.1 Extraction terminologique

Un modèle terminologique qui serait basé exclusivement sur des mots simples ne serait pas assez riche. Ainsi, les expressions composées participant à l'indexation sont à la base de la terminologie d'un domaine. On envisage alors l'extraction des termes du domaine à l'exclusion du simple stockage des termes-index inscrit dans les notices.

L'algorithme que nous utilisons est le suivant. Lorsqu'un mot courant lu dans le texte est reconnu comme forme fléchie grâce au dictionnaire, alors sa forme lemmatisée est récupérée. Dans le cas où rien ne serait trouvé, on opère une troncature du suffixe si un suffixe connu est identifié [Porter, 1980]. De cette façon on obtient en sortie la forme canonique ou racine de la forme brute issue du texte. Quatre types de méthode conduisent alors à l'extraction de groupes nominaux: l'application d'un patron morpho-syntaxique, l'utilisation d'un dictionnaire, la méthode des segments répétés ou la méthode des bornes.

Nous utilisons ici la méthode des bornes qui est simple à implémenter et à adapter, mais qui s'avère robuste par rapport à la variété des documents et des langues naturelles susceptibles d'être rencontrés [Turenne, 2000]. Nous décrivons la méthode comme suit. On associe une expression à la détection du Groupe Nominal (GN): B M* B, où B est une borne et M* est le GN, c'est-à-dire une suite de mots de taille quelconque. Un automate parcourt le texte du début à la fin. Il découpe morphologiquement chaque mot et l'identifie à un mot contenu dans un dictionnaire. S'il trouve une correspondance alors ce mot est interprété comme une borne. Il enregistre comme groupe nominal tout ce qui suit après la borne jusqu'à rencontrer une nouvelle borne. Il stocke le groupe nominal courant dans une liste et continue ainsi jusqu'à la fin du texte.

4.2 Extraction de classes

Dans l'analyse d'un réseau d'auteurs deux informations sont récupérables d'un auteur par rapport aux autres: les cliques de co-auteurs auxquelles il appartient et les auteurs qu'il cite.

Un premier algorithme va construire des classes avec les co-auteurs. L'algorithme utilise les propriétés séquentielles des données pour extraire des classes d'auteurs évoluant au cours du temps. La méthode est basée sur l'utilisation de l'algorithme *Apriori* [Agrawal et Srikant,

1994], assimilable à un algorithme de recherche de cliques. Un deuxième algorithme utilise les citations des auteurs dans leurs publications. On peut ainsi construire des classes d'auteurs évoluant au cours du temps partageant les mêmes auteurs participant au référentiel de leurs travaux. L'algorithme ne développe le treillis qu'aux niveaux de base, ce qui permet d'une part de réduire le nombre de classes et d'autre part de calculer les classes en un temps raisonnable.

Les deux algorithmes se complètent dans la mesure où les classes obtenues par construction de motifs sont de petite taille contrairement à celles obtenues par le treillis de concepts. Le facteur d'échelle est différent tout comme leur nature. En effet dans le cas des motifs d'auteurs, la classe s'interprète comme un travail de collaboration entre auteurs alors que les classes du treillis correspondent à des identités plutôt d'appartenance communautaire.

4.3 Indices de la dynamique de production des connaissances

Les algorithmes précédents ont pour rôle de fournir des ensembles de descripteurs pour lesquels il s'agit ensuite de conduire l'étude de leur variation temporelle grâce à une visualisation des motifs proposée par Beluga. On peut ainsi tester des hypothèses sur la variation des motifs de co-publication par exemple. Si de telles quantités ne se manifestent pas de manière évidente et n'induisent pas directement des schémas interprétables, alors des techniques de visualisation sont nécessaires. Evidemment plusieurs solutions sont possibles à ce niveau, et notamment la distribution des valeurs de ratios ou d'indices en fonction du temps.

Nous avons développé une adaptation de la méthode du Graphe Socio-Technique [Latour et al., 1991] pour établir des indices de la dynamique des connaissances certifiées d'un domaine scientifique. Cette méthode propose initialement des indices pour cerner le double problème d'obtenir des formes de quantification adaptées au caractère toujours circonstanciel des processus d'innovation et de comparer des processus d'innovation entre eux.

On définit alors les indices élémentaires suivants basés essentiellement sur un comptage des ensembles de descripteurs d'une période donnée par rapport à la précédente ou aux précédentes : N_t caractérise le nombre de *nouveaux descripteurs en t par rapport à t-1* et I_t caractérise celui des *descripteurs invariants de t-1 à t,* T_t étant le nombre de *descripteurs à t.* Ces indices permettent de proposer ensuite des indices : l'*indice de changement* $IC_t = N_t / T_t$ qui explique l'intensité des changements, l'*indice de nouveauté* $IN_t = N_t / \Sigma Ni$ qui marque l'intensité relative des changements et l'*indice de persistance*, $IP_t = I_t / T_{t-1}$.

Ces indices permettent alors de caractériser des trajectoires d'évolution de la composition des champs. Un exemple est donné ci-dessous pour la description du champ " Auteur " entre 1964 et 2002 (FIG.1). La croissance rapide des indices IP et IN après 1996 corrobore bien la dynamique de croissance et de renouvellement du domaine suite à l'épizootie de l'Esb en Europe continentale. On notera aussi la chute de IC entre 1990 et 1993, période de grande fragilité pour le financement des recherches sur une théorie alors récente : celle du prion.

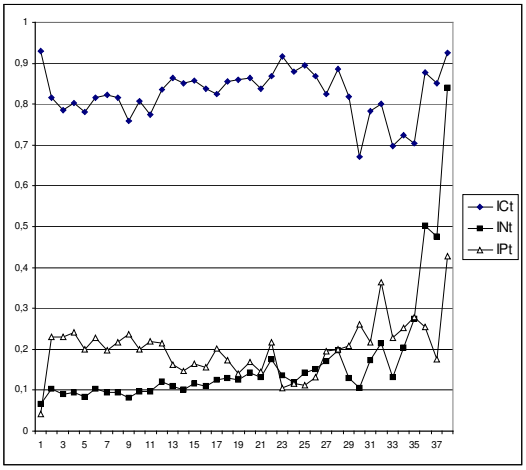

FIG.1 – *Indice : Corpus ESST/Prions. Période de 1 an - Champ des Auteurs*

5. Conclusion

Dans cet article nous avons présenté un système de traitement de l'information basé sur la recherche et l'extraction d'information textuelle. Le système reflète une collaboration étroite joignant une vision du traitement documentaire propre à la sociologie et une vision propre à l'informatique. Les enjeux d'une telle interaction pluridisciplinaire sont multiples. En informatique avec les moyens de stockage et de traitement de l'information, on comprend aisément les nouveaux enjeux représentés par le document électronique: concevoir de nouvelles méthodes d'accès à l'information et faciliter le classement, la recherche, la synthèse et la diffusion.

Dans ce cadre, le domaine informatique de la fouille de textes est prédisposé à offrir des techniques d'acquisition de connaissances et des techniques de visualisation de la dynamique des connaissances. Cet objectif nous a permis de développer la plate forme Beluga pour viser le traitement de l'information textuelle et d'extraction et de visualisation de connaissances. L'outil est fortement inspiré des moteurs d'indexation connus en informatique documentaire mais implémente des fonctionnalités d'extraction d'information proposées à l'utilisateur. Par son ouverture il nous permet d'envisager d'y intégrer de nouvelles fonctionnalités et en particulier offre la possibilité de dérouler les traitements sur d'autres domaines scientifique c'est-à-dire d'autres corpus de textes et d'autres problématiques de recherche.

Références

[Agrawal et Srikant, 1994] R. Agrawal et R. Srikant. Fast Algorithms for Mining Association Rules, *Proc. of the 20th Int. Conf. on Very Large Databases*, Santiago, Chile, 1994.

[Barbier et De Looze, 1999] M. Barbier et M-A. De Looze. A scientometric description of the evolution of the TSEs field. Networks of authors and research themes in the MEDLINE Database", *International Symposium on Characterization and Diagnosis of Prion Diseases in Animals and Man*, Poster, Tübingen, 1999.

[Bar-Ilan, 1997] J. Bar-Ilan. The Mad Cow Disease, Usenet Newsgroups and Bibliometric Laws. *Scientometrics*, 39 (1): 29-55, 1997.

[Benzécri, 1973] J-P. Benzécri. L'analyse des données II. La taxinomie. Dunod, Paris, 1973.

[Chateauraynaud et Torny, 1999] F. Chateauraynaud et D. Torny. Les sombres précurseurs. Une sociologie pragmatique de l'alerte et du risque. Editions de l'EHESS, Paris, 1999.

[Chen et al., 2002] C. Chen, T. Cribbin, R. Macredie et S. Morar. Visualizing and Tracking the Growth of Competing Paradigms: Two Case Studies. *Journal of the American Society For Information Science and Technology,* 53(8): 678–689, 2002.

[Ganter et Wille, 1999] B. Ganter et R. Wille. Formal Concept Analysis: Mathematical Foundations Springer, Berlin.

[Latour et al., 1991] B. Latour, P. Mauguin et G. Teil. Une méthode nouvelle de suivi socio-technique des innovations: le graphe socio-technique, *in* D. Vinck (sld) Gestion de la recherche. Nouveaux problèmes, nouveaux outils. Armand Colin, Paris, 1991.

[Lelu et Ferhan, 1998] A. Lelu et S. Ferhan, Clustering a textual dataflow by incremental density-modes seeking. *Proc of International Federation of Classification Societies,* Université La Sapienza, Rome, 1998.

[Li et al., 2002] Y. Li, S. Zhu, X. Wang et S. Jajodia. Looking into the seeds of Time: Discovering Temporal Patterns in Large Transactions Sets, Technical report. Ed. George Mason University, Arlington, 2002.

[Maunoury et al., 1999] M-T. Maunoury, A-M. De Recondo et W-A.Turner. Observer la science en action. *Médecine/Sciences*, 15 (4): 577-582, 1999.

[Porter, 1980] M-F. Porter. An Algorithm for Suffix Stripping. *Program*, (14): 130, 1980.

[Stone et Kelly, 1966] P-J. Stone et E. Kelly. *The General Inquirer: A Computer Approach to Content Analysis*, The MIT Press, Cambridge, MA, 1966.

[Turenne, 2000] N. Turenne. *Apprentissage statistique pour l'extraction de concepts à partir de textes. Application au filtrage d'informations textuelles*. Thèse de doctorat, Université Louis Pasteur - INSA de Strasbourg, 2000.

[White et Griffith, 1981] H-D. White et B. Griffith. Author co-citation: A literature measure of intellectual structure. *Journal of the American Society for Information Science*, 32:163-171, 1981.

Summary

Thanks to an interdisciplinary project on Prion diseases a common methodology of thematic analysis, based on applied computer science and social studies of science, is set up in order to understand the dynamic of scientific knowledge about these diseases. We developed a platform named ‘Beluga’ for scientific notices indexing which aim is to enable the extraction of associations changing over periods of time. Beluga is proposing some modules based on documents indexing according to units of analysis: references, authors, terms, institutions. This tool is based on a double approach of learning and visualisation that automates the process of extracting groups of authors or terms and allows the user to get back to documents. This diachronic analysis of electronic documents corpora enables to describe how the various themes are structuring their terminology while leading to consider the problems of emerging thematic features.