Ingénierie de l'Information Scientifique et Technique : Fouille de Données Textuelles

Ivana Roche, Claire François

Ingénierie de l'Information Scientifique et Technique : Fouille de Données Textuelles [1]

Ivana ROCHE[2]
Claire FRANCOIS[2]
Institut de l'Information Scientifique et Technique du
Centre National de la Recherche Scientifique, França

RESUME

Nous proposons de présenter l'approche de fouille de données textuelles développée dans notre équipe. Nos travaux s'inscrivent dans une démarche d'analyse de l'information scientifique et technique, afin d'aider les utilisateurs à enrichir leur modèle de connaissances ou à effectuer une tâche complexe comme la veille technologique, le traitement de données scientifiques, ou la prise de décision stratégique dans la gestion scientifique. Aussi nos objectifs sont de construire une représentation de l'information présente dans les documents et de capitaliser les connaissances acquises en rendant cette représentation plus compréhensible par les utilisateurs. Pour répondre à ces objectifs nous avons développé un processus de fouille de textes qui réalise une représentation de l'information présente dans les documents en intégrant des techniques de traitement automatique des langues et des mathématiques appliquées à l'analyse de l'information scientifique. Dans les différentes étapes du traitement des textes la prise en compte des connaissances du domaine est nécessaire afin de pouvoir analyser et raisonner sur le contenu des documents. Pour outiller cette approche nous avons réalisé une station d'analyse de l'information permettant la recherche d'information dans des bases de données documentaires en vue de leur analyse statistique, terminologique, thématique et cartographique.

MOTS CLES

analyse de l'information; représentation de l'information; information scientifique et technique; modèle de connaissances; fouille de données textuelles.

TEXTE DU TRAVAIL

Introduction

Le but de nos travaux est la conception de méthodes et outils pour l'analyse de l'information des textes scientifiques et techniques, en appliquant des méthodes de type

[1] Trabalho apresentado no IX Colóquio Brasil-França de Ciências da Comunicação, evento componente do XXXII Congresso Brasileiro de Ciências da Comunicação

[2] Ingénieur R&D à l'INIST / CNRS, emails : [ivana.roche ; claire.francois]@inist.fr

classification afin de construire une représentation structurée de leur contenu. Pour un expert du domaine scientifique traité, cette représentation est une aide à l'analyse. A terme, nous espérons pouvoir capitaliser les connaissances acquises sous une forme exploitable pour l'analyse de nouveaux corpus.

Nos travaux sont orientés vers des applications en veille scientifique et technologique et des études de la science et de la technologie afin de répondre à des besoins de stratégiques de gestion de la science.

De façon plus précise, nous désirons pouvoir répondre à des questions portant, d'une part, sur le contenu scientifique des publications et, d'autre part, sur la relation entre les acteurs de la recherche et le contenu de leurs recherches, par exemple, étant donné un domaine scientifique :

- ► Quelles sont les thématiques qui se développent actuellement, ou au contraire sont en déclin ?
- ► Quels sont les auteurs et les laboratoires leaders dans ces thématiques ?
- ► Quelles relations ces auteurs entretiennent-ils entre eux et avec les auteurs dans d'autres thématiques ?
- ► Où publient-ils préférentiellement leurs travaux ?

Nos objectifs sont donc :

- construire une représentation de l'information présente dans les documents,
- rendre cette représentation plus compréhensible par les utilisateurs,
- capitaliser les connaissances acquises.

Le processus de fouille de textes mis en place construit une représentation de l'information présente dans les documents en intégrant les techniques de traitement automatique des langues, l'exploitation de la structure interne de documents, et des outils de mathématiques appliquées à l'analyse de l'information scientifique (classification, cartographie). Dans les différentes étapes du traitement des textes la prise en compte des connaissances du domaine est nécessaire afin de pouvoir analyser et raisonner sur le contenu des documents. Ces connaissances existent à différents niveaux de structuration (terminologies, bases de données factuelles, bases de connaissances, ontologies).

Démarche scientifique

La démarche d'analyse de l'information que nous avons défini jusqu'à présent est représentée dans la figure 1.

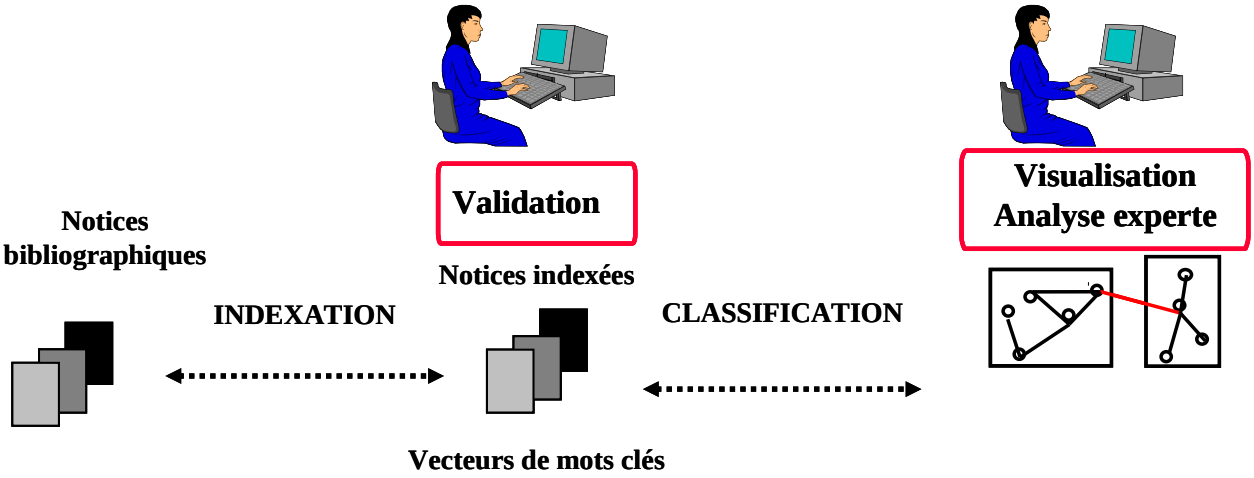

Figure 1 : Schéma global du processus d'analyse défini et mis en oeuvre

Ce processus est itératif et interactif (assisté par ordinateur). Il correspond au cas général de classifications calculées à partir des mots-clés, cependant nous avons également appliqué cette démarche occasionnellement à partir d'autres champs comme les auteurs, ou les liens entre sites Web. Nous détaillons ses étapes dans le tableau 1.

Etape	Description
1	Définition du corpus d'étude
2	Application de statistiques descriptives pour obtenir une première vue d'ensemble
3	Utilisation de l'indexation native ou d'une indexation assistée, structurée ou non selon les ressources disponibles
4	Validation manuelle de l'indexation fondée sur les critères sémantiques définis par l'expert ou sur une classification syntaxique hors contexte (pour l'indexation assistée)
5	Classification automatique et cartographie
6	Analyse manuelle des classifications

Tableau1 : Description des phases du processus d'analyse défini et mis en oeuvre

Indexation assistée

Pour indexer les documents ou notices bibliographiques, nous avons développé la ***PLATE-FORME ILC*** (Infométrie, Langage, Connaissance) (ROYAUTE 1999, JACQUEMIN et al., 2002), dont l'objectif est l'indexation automatique contrôlée de documents à partir de ressources terminologiques. Cet outil permet de retrouver les termes des ressources sous leur forme normale ou sous une forme variante. Trois catégories de variations sont traitées : (a) la variation flexionnelle, (b) la variation syntaxique et (c) la variation morpho-dérivationnelle.

La seconde spécificité de cet outil est qu'il permet de définir la ressource terminologique qui sera utilisée pour l'indexation, et de la prétraiter de la même manière que les textes à indexer. Cet outil est donc plus ouvert que la plupart des outils d'indexation qui sont fournis avec leurs dictionnaires précompilés. Il permet de définir une indexation cohérente à partir de documents de diverses provenances.

Classification automatique et cartographie

Nous avons développé deux outils de classification et de cartographie.

L'outil SDOC est une réalisation informatique de la méthode des mots associés (CALLON et al., 1983, GRIVEL et al. 1995)dont l'objectif est de mettre en évidence la structure de leurs relations en se fondant sur leur cooccurrence. La notion de cooccurrence est essentielle car si on considère que deux documents sont proches parce qu'ils sont indexés par des mots-clés similaires, alors deux mots-clés figurant ensemble dans un grand nombre de documents seront considérés comme proches. L'emploi d'un indice statistique permet de normaliser la mesure de l'association entre deux mots-clés. Après le calcul du réseau des associations des mots-clés, SDOC applique un algorithme de classification ascendante hiérarchique (CAH), dit du simple lien (« single link clustering »), afin de construire des classes de mots proches les uns des autres. Les classes sont ensuite positionnées sur un plan cartésien selon les valeurs de leurs « densité » et « centralité », constituant ainsi une carte.

A partir d'une représentation vectorielle des documents, l'outil Neurodoc classe les documents à l'aide d'une méthode de classification non hiérarchique, les K-means axiales (LELU 1990, LELU & FRANCOIS 1992), puis calcule représentation des classes obtenues sur une carte à l'aide d'une analyse en composantes principales (ACP). Les classes ainsi obtenues sont des indicateurs des thèmes ou des centres d'intérêt

autour desquels s'agrège l'information, tandis que la carte propose une visualisation globale des thèmes et représente un indicateur stratégique permettant d'apprécier la position relative des classes dans l'espace de connaissance.

Station d'analyse de l'information : Stanalyst

Nous avons regroupé ces outils dans STANALYST (POLANCO et al., 2001), une station d'analyse de l'information permettant la recherche d'information dans des bases de données documentaires de l'INIST / CNRS en vue de leur analyse. L'organisation de la station est présentée dans la figure 2.

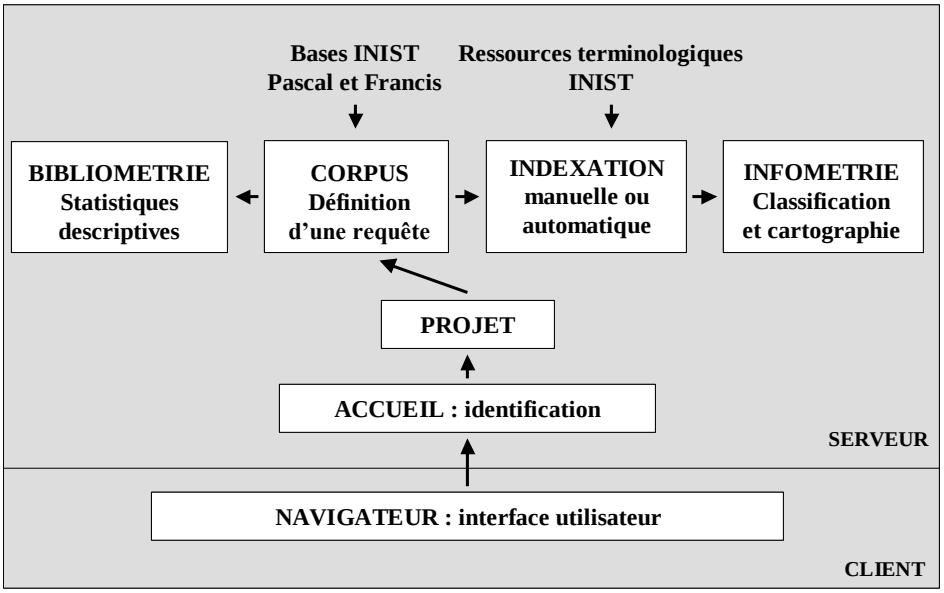

Figure 2 : Organisation de la station d'analyse de l'information STANALYST

Cette station d'analyse, accessible depuis un navigateur web, intègre sous une interface commune un ensemble de modules applicatifs permettant la recherche d'information dans les bases de données bibliographiques de l'INIST, en vue de leur analyse statistique, terminologique, thématique et cartographique.

Les caractéristiques du système sont les suivantes :

- utilisation d'un serveur HTTP Apache,
- programmes CGI développés en Perl pour l'interface Web utilisable depuis un navigateur,
- programmes en différents langages pour les traitements,

- données et index en SGML manipulés à l'aide d'une bibliothèque de commandes développée à l'INIST-CNRS et stockés dans un système de fichiers.

Pour l'instant, STANALYST fonctionne avec le format interne de l'INIST / CNRS utilisé pour ses bases bibliographiques PASCAL et FRANCIS.

Chaque module de la station est le reflet d'une fonctionnalité bien précise :

- le module PROJET gère la création de projets. Chaque projet est un répertoire dans lequel seront stockés tous les résultats le concernant et définissant ainsi un environnement de travail. L'utilisateur en est le propriétaire, mais il a également la possibilité de donner accès à son projet aux utilisateurs associés reconnus par la station ;
- le module CORPUS gère la création de corpus, à partir d'un outil qui permet la génération de corpus par exécution de requêtes construites par l'utilisateur. Les corpus peuvent ensuite être exportés à destination des modules suivants ;
- le module BIBLIOMETRIE permet à l'utilisateur de choisir et de produire des analyses statistiques correspondant aux indicateurs bibliométriques classiques ;
- le module INDEXATION permet de réviser l'indexation préexistante dans les notices bibliographiques en vue d'une classification thématique. Grace à l'outil ILC, il permet également de réaliser une indexation automatique du corpus à l'aide de plusieurs référentiels terminologiques intégrés à STANALYST ®. Il permet également la visualisation et la validation, par l'utilisateur, des résultats ;
- le module INFOMETRIE gère la classification thématique à partir des deux outils SDOC et Neurodoc.

Dans le but de gérer le transfert de données d'un module à l'autre, les données des utilisateurs sont organisées en répertoires sur le serveur sur un mode arborescent partant depuis les répertoires PROJET. L'ensemble des traitements se fait au sein de l'arborescence d'un même projet reflétant ainsi l'enchaînement logique de la démarche d'analyse de l'information.

Nous ouvrons cette station aux utilisateurs désireux de réaliser eux-mêmes leurs analyses. L'ouverture d'un compte utilisateur protégé par un mot de passe se fait sur simple demande. A ce jour, STANALYST compte plus de 200 utilisateurs dont environ 60 % extérieurs à notre institut. Parmi ces derniers, près de 15 % sont localisés à

l'étranger. A titre indicatif, au cours du premier semestre 2009 près de 9000 requêtes ont été exécutées.

Dans un souci d'ouverture de la station de travail à d'autres sources d'information scientifique et technique nous avons, dans le cadre d'un projet financé par le Ministère des Affaires Etrangères français, collaboré avec le BIREME[3] (Brésil), le CAICYT[4] (Argentine), la CONICYT[5] (Chile) et la RICYT[6] (Argentine) afin de développer une version de STANALYST (POLANCO et al., 2007) capable d'exploiter, dans un premier temps, les bases SciELO[7] mais aussi d'autres bases (LILACS, MEDLINE, bases de brevets, bases régionales).

Les développements réalisés dans le cadre de ce projet ont, parallèlement à l'ntroduction du traitement multibases, permis d'avancer sur deux autres points importants :

- le portage de la station sous LINUX,
- l'internationalisation de l'interface, désormais disponible également en espagnol, portugais et anglais.

D'autre part, les références bibliographiques des bases SciELO donnant accès aux documents électroniques associés, la mise à disposition d'un tel réservoir d'information peut également être considérée comme une retombée du projet mais s'inscrivant à plus long terme, dans le cadre de l'évolution de la station STANALYST® vers le traitement d'unités d'information en texte intégral.

Actuellement, une nouvelle version de la station, tournant sous LINUX et comportant un nombre significatif d'avancées, est en cours de test chez un groupe d'utilisateurs internes.

[3] Centro Latino-Americano e do Caribe de Informação em Ciências da Saúde

[4] Centro Argentino de Información Científica y Tecnológica

[5] Comisión Nacional de Investigación Científica Y Tecnológica

[6] Red Iberoamericana/Interamericana de Indicadores de Ciencia y Tecnología

[7] SciELO est une initiative plurinationale visant à réunir et diffuser sous forme électronique la littérature scientifique publiée dans les pays d'Amérique Latine et des Caraïbes, l'Espagne et le Portugal par le développement d'une méthodologie commune de traitement, stockage, diffusion et évaluation de cette littérature (SciEntific Library Online - http://www.scielo.org/)

Positionnement et développements en cours

Notre objectif est maintenant de compléter l'approche analytique actuelle en dépassant les limites que nous avons observées et en nous rapprochant du schéma défini dans la figure 3.

Ainsi, nous avons vu que les documents traités sont encore très majoritairement des notices bibliographiques (figure 1). Ces documents respectent une structure prédéfinie fixe, avec des champs bien identifiés contenant des informations catalographiques telles les noms des auteurs et leurs affiliations, le titre, le millésime, etc, et dans lesquels la partie textuelle la plus volumineuse que l'on puisse trouver est un simple résumé. Or, ne traiter que ce type d'information restreint le nombre de sources d'information exploitables, nous imposant l'utilisation quasi exclusive de bases de données bibliographiques ou de brevets ce qui limite de manière non négligeable l'analyse que nous pouvons réaliser. Les documents complets électroniques sont de plus en plus disponibles, et « les diffuser et les signaler » est d'ailleurs une tendance qui s'accentue et une orientation importante de l'INIST. Savoir traiter ces documents en tirant parti de la richesse de l'information disponible pose le problème de leur représentation. En effet, il faut pouvoir traiter les documents complets avec comme double objectif d'exploiter les informations exprimées en langage naturel dans les documents, d'une part, et les informations contenues dans la structure même de ces documents, d'autre part.

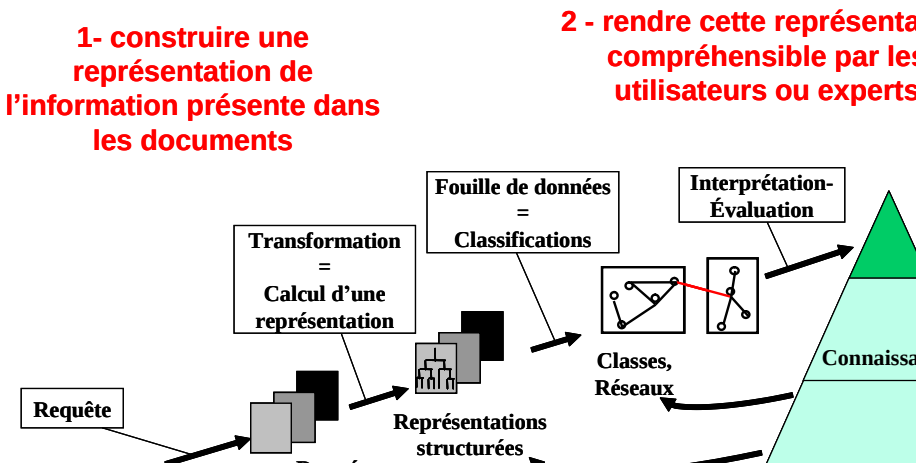

Figure 3 : Processus de fouille de texte et objectifs de recherche

Dans la perspective de travailler sur des représentations structurées des documents, il est également important de faire évoluer nos méthodes de classification afin de prendre en compte ces structures. De plus, si au terme d'une analyse complète nous obtenons une image du corpus ou d'un domaine scientifique, cet image reste figée à un instant donné alors que l'étude de l'évolution et de la dynamique d'un domaine est nécessaire. Ces limites doivent pouvoir être dépassées avec le développement de méthodes de classification de documents portant, d'une part, sur la prise en compte de la structure des données en entrée et, d'autre part, sur le suivi temporel de l'évolution de l'information afin repérer les thématiques émergentes ou bien des redéfinitions de thématiques en sous thématiques.

Enfin, nous avons observé que les étapes 4 et 6 du processus d'analyse (tableau 1) sont des grandes consommatrices en temps-expert. En effet, l'interprétation des classifications reste à la charge de l'expert qui doit confronter lui-même les structures mises en évidence avec ses connaissances du domaine. De plus, ce travail n'est pas capitalisé en vue d'une utilisation ultérieure lors d'une analyse sur un même domaine. L'objectif est donc de compléter notre processus avec une aide à la capitalisation des connaissances acquises.

Nous positionnerons donc résolument nos travaux dans le domaine de la fouille de textes : comme la fouille de données a pour objet de d'extraire des informations utiles des bases de données en utilisant des techniques d'analyses de données, la fouille de textes s'applique à extraire des informations utiles, non triviales et implicites à partir de textes. Par conséquent, nos travaux viendront se situer à l'intersection de différents champs disciplinaires :

- TAL[8] et traitement des documents structurés,
- classification automatique
- extraction, gestion et capitalisation de connaissances

[8] Traitement Automatique de Langues

TAL et traitement des documents structurés

Nos travaux étudieront comment définir et construire une représentation des documents tenant compte à la fois de leur contenu et de leur structure. Dans ce cadre, nous participons déjà à deux projets : le projet Rhecitas et le projet QUAERO

Le projet Rhecitas[9] constitue un aspect très ciblé de la problématique de construction d'une représentation de l'information présente dans les documents. L'objectif de ce projet est de réaliser un prototype d'annotation des liens bibliographiques sur un corpus d'articles en français et dans le domaine des SHS. L'originalité du projet porte sur l'identification la fonction rhétorique des relations de citations, c'est-à-dire la catégorisation de ces dernières en fonction du rôle que l'auteur leur attribue (par exemple filiation des approches, désaccord, références de type méthodologique, définition, …). Ces travaux utilisent des méthodes d'analyse automatique de texte et sont basés sur un modèle existant, comportant 12 fonctions, dont la caractérisation a été proposée pour l'anglais par Simone Teufel[10].

Dans le projet QUAERO[11], nous travaillons sur l'annotation sémantique de documents scientifiques et techniques. Cette annotation correspond à une forme très élaborée de la représentation de l'information présente dans les textes. Notre apport dans ce projet est une réflexion sur l'annotation manuelle des documents nécessaire à l'apprentissage et à l'évaluation des outils développés pour l'annotation automatique, ainsi que sur l'acquisition de ressources permettant de représenter des connaissances élaborées de type ontologies.

Classification automatique

Les méthodes de classification et de cartographie sont particulièrement adaptées à la structuration d'une grande masse d'information. Après avoir développé deux méthodes déjà intégrées dans la plateforme STANALYST®, notre problématique est d'étudier les évolutions temporelles afin de détecter des évolutions et émergences.

[9] Projet financé par le TGE ADONIS, appel à projet 2007, http://www.tge-adonis.fr/?Le-projet-RHECITAS-RHEtorique-des
[10] voir le site du projet CitRAZ : http://www.cl.cam.ac.uk/~sht25/Project_Index/Citraz_Index.html, et plus précisément ses articles de 2006 disponibles sur cette page.
[11] http://www.quaero.org/modules/movie/scenes/home/

Aussi dans le cadre du **projet européen PROMTECH**[12]**,** nous avons développé une méthodologie d'identification de technologies émergeantes et prometteuses ayant des liens forts avec la Physique et des communautés scientifiques les développant (SCHIEBEL et al., 2009 ; BESAGNI et al., 2009 ; ROCHE et al., 2008 ; ROCHE et al., 2007). L'objectif principal de ce projet a été de fournir un outil d'aide à la sélection de thématiques qui ont été soutenues dans les programmes de R&D du 7ème Programme-Cadre de la Commission Européenne.

Pour traiter ce problème non résolu de suivi de l'évolution temporelle de l'information, nous animons un projet placé dans le cadre du Contrat de Projet Etat Région (CPER) Lorraine 2007-2013[13] : **le projet McFIID** (Multiclustering de flux d'informations incrémental et distribué). Son objectif est de développer des outils de veille prenant en compte le facteur temps afin de suivre l'évolution des thématiques de recherche d'un domaine scientifique donné en termes d'émergence ou de déclin. Notre choix se porte sur le développement de méthodes de classification non supervisée incrémentales qui permettent de détecter de convergences et divergences de thématiques à partir de corpus de documents (CUXAC et al., 2009). Dans ce projet, nous tirons parti du caractère intrinsèquement distribué de nos algorithmes pour les développer sur une architecture répartie à base de Web-services. Cette architecture permet également de réaliser plusieurs classifications sur différentes facettes des données et de les croiser afin de réaliser une analyse complète de l'information.

Extraction, gestion et capitalisation de connaissances

La combinaison classification – analyse symbolique par règles est une approche qui commence tout juste à se développer, cela s'inscrit dans la démarche plus générale de la combinaison de méthodes symboliques et numériques.

Par ailleurs, nous poursuivons l'objectif de consolider notre positionnement au sein de la communauté infométrie /scientométrie avec des actions comme notre participation au **projet européen DBF**[14]. Ce projet s'inscrit dans le cadre des actions de coordination et

[12] Le projet PROMTECH (PROMising TECHnologies) est arrivé à terme en septembre 2007. Il a été réalisé dans le cadre de l'Action Spécifique NEST (New and Emerging Science and Technology - http://www.cordis.lu/nest/whatis.htm) du 6ème Programme-Cadre de Recherche et Développement de l'Union Européenne. Le Consortium PROMTECH a réuni l'ARC System research GmbH (Vienne, Autriche), le Fraunhofer Institut für Systemtechnik und Innovationsforschung (Karlsruhe, Allemagne) et l'INIST/CNRS.

[13] http://talc.loria.fr/

[14] Development and Verification of a Bibliometric Model for the Identification of Frontier Research

appui du Programme IDEAS du 7ème PCRD[15] placé sous la responsabilité du European Research Council (ERC). Il démarrera en septembre 2009 et aura une durée de trois ans. Nous rappelons que la principale mission du ERC est le financement de projets de recherche menés par des chercheurs européens. En 2007, un peu plus de 9000 soumissions ont été enregistrées et environ 300 ont été sélectionnées. Lors du processus de sélection, l'ERC déploie un réseau d'experts chargés d'examiner l'ensemble des propositions, de les analyser et de faire un choix fondé sur des critères énoncés en fonction de la notion de recherche « frontière » que l'on situe au premier rang du processus de création de nouvelles connaissances.

Le projet DBF développe une problématique d'analyse qui a pour objectif d'apporter aux experts en charge de la sélection des projets une information complémentaire et objective sur les choix opérés sous la forme d'indicateurs. La méthodologie que nous proposons consiste en :

- modéliser le processus de sélection des propositions de projets reçues à l'aide d'indicateurs bibliométriques/scientométriques. La modélisation doit reproduire les critères utilisés dans le processus de sélection ERC : innovation, facteur de risque, interaction fondamental/appliqué et interdisciplinarité,
- appliquer le modèle sur les données collectées lors d'un appel ERC,
- comparer a posteriori les résultats obtenus avec les résultats annoncés par les experts ERC,
- assurer un feedback sur, à la fois, les experts ERC et notre modèle.

Conclusion

Les perspectives de recherche pour les prochaines années au sein de notre équipe sont très ouvertes et plusieurs défis se profilent devant nous.

Le premier défi porte sur l'exploitation de grandes masses de données numérisées. Ces données sont de type Information Scientifique et Technique (IST) provenant de grandes bases de données documentaires, sites Web académiques et des sites de publications en ligne. Pour ce type de données, nous poursuivrons notre démarche d'analyse du contenu (quelles sont les thématiques de recherche et comment s'organisent-elles ?), nous la

[15] Programme-Cadre de Recherche et Développement

couplerons avec l'analyse du réseau social associé qui correspond aux producteurs de ces informations.

Le second défi porte sur le traitement des données :

- Comment rendre ces grandes masses de données intelligibles ?
- Comment en extraire une information utile ?

Pour cela, nous proposons d'intégrer différentes approches comme les méthodes statistiques : classification, cartographie, théorie des graphes, les méthodes de fouille de données et de texte les méthodes de traitement automatique des langues (TAL) et enfin les méthodes de visualisation de l'information.

Dans le monde de la science et de la technologie, nous pouvons observer l'omniprésence du document numérique de l'acquisition des données brutes à la publication des résultats associées à l'émergence de différents types d'infrastructures de production et de partage de données, à différentes échelles. Il devient donc nécessaire de croiser ces différents types de données et d'être capable de gérer cette hétérogénéité. Ce phénomène associé au développement des pratiques collaboratives sur le Web induit également des changements de comportement des acteurs. Nous avons donc débuté un axe de recherche portant sur l'analyse de ces appropriations par les chercheurs de ces technologies numériques dans le cadre de l'étude d'un processus d'écriture collective dans un environnement de type Web 2.0 (**projet INSI**[16])

Aussi notre problématique de recherche, déjà amorcée, continuera à se développer principalement autour de la définition : d'environnements d'exploration intégrant les dernières avancées dans les domaines du TAL et de l'analyse de données (indicateurs, classification, graphes et cartographie), ainsi que des outils d'extraction et de gestion de connaissances. Cette problématique devra intégrer la diversification du type de données à traiter mais également l'évolution des pratiques et des technologies disponibles.

REFERENCES

SCHIEBEL E., HÖRLESBERGER M., ROCHE I., FRANCOIS C., BESAGNI D., Identifying and Characterising Technological Topics in the Field of Optoelectronic Devices: Two Complementary Methodological Approaches, 12th International Conference on Scientometrics and Informetrics, Rio de Janeiro, 14-17 July 2009

[16] Projet « Atelier d'usage des informations numériques pour la science et l'innovation (INSI) » financé par l'Institut des Sciences de la Communication du CNRS - http://maquettewicri.loria.fr/fr.incubWicri/index.php5?title=Ateliers_d%27usage_INSI

BESAGNI D., FRANCOIS C., HÖRLESBERGER M., ROCHE I., SCHIEBEL E., Les émergences technologiques dans le domaine des dispositifs optoélectroniques : identification et caractérisation, 2nd Séminaire VSST - Veille Stratégique Scientifique & Technologique, Nancy, 30-31 mars 2009

CALLON M., COURTIAL J-P., TURNER W.A., BAUIN S. From Translation to Problematic Networks: An Introduction to Co-Word Analysis; Social Science Information, vol. 22, pp. 191-235; 1983

CUXAC P., LELU A., CADOT M. : Suivi incrémental des évolutions dans une base d'information indexée : une boucle évaluation / correction pour le choix des algorithmes et des paramètres. Systèmes d'Information et Intelligence Economique (SIIE'09), 12-14 Février 2009, Hammamet, Tunisie.

GRIVEL L., MUTSCHKE P., POLANCO X. (1995) "Thematic mapping on bibliographic databases by cluster analysis : a description of the SDOC environment with SOLIS", *Journal of Knowledge Organization*, Vol. 22, 1995, n° 2, p. 70-77

JACQUEMIN C., DAILLE B, ROYAUTE J. et POLANCO X., 2002 - In Vitro Evaluation of a Program for Machine-Aided, Information Processing & Management, Vol. 38, No. 6, pp. 765-792

LELU A. ; Modèles neuronaux pour données textuelles - Vers l'analyse dynamique des donnée; Journées ASU de Statistiques, Tours ; 1990

LELU A., FRANCOIS C. ; Automatic generation of hypertext links in information retrieval systems; Colloque ECHT'92 ; Milan, D. Lucarella & al. Eds., ACM Press, New York ; 1992

ROCHE I., BESAGNI D., FRANCOIS C., HÖRLESBERGER M., SCHIEBEL E., Identification and Characterisation of Technological Topics in the Field of Molecular Biology, 10th International Conference on Science and Technology Indicators, Vienna, 17-20 September 2008 (extended version to be published in Scientometrics)

ROCHE I., FRANCOIS C., BESAGNI D., Détection de techniques prometteuses à partir de méthodes bibliométriques, 10ème Colloque International sur le Document Electronique, Nancy, 2-4 juillet 2007

POLANCO X., ALBORNOZ M., PACKER A., PRAT A. M., BESAGNI D., FRANCOIS C., ROCHE I., BARRERE R., MATAS L. J., SANTOS F. B. C. dos, WALTERS J., STANALYST-SciELO: Modelo y uso para la vigilancia científica, VII Congresso Iberoamericano de Indicadores de Ciencia y Tecnologia, São Paulo, 23-25 de maio de 2007

POLANCO X., FRANCOIS C., ROYAUTE J., BESAGNI D., ROCHE I., Stanalyst®: An integrated environment for clustering and mapping analysis on science and technology. In: Proceedings of the 8th International Conference on Scientometrics and Informetrics, Sydney, 16-20 July 2001

ROYAUTE J., 1999 - Les groupes nominaux complexes et leurs propriétés : application à l'analyse de l'information, Thèse de doctorat en informatique, Université Henri Poincaré - Nancy I, 19 juillet 1999, 228 pages

FAAYAD U., PIATETSKY-SHAPIRO G., SMYTH P., UTHURUSAMY R., Eds., From Data Mining to Knowledge Discovery, Chapter 1, 1996