CAIS ACSI

Canadian Association for Information Science
L'Association canadienne des sciences de l'information

Home
About CAIS
CAIS Executive
Membership
Journal
Conferences
Contact CAIS
Links
Site map

Français

Search

CAIS Conferences

For more than a quarter of a century, Canadian information scientists and professionals have met to discuss the *access*, *retrieval*, *production*, *value*, *use*, and *management of information*. From those early days of examining computational ways of manipulating information through to investigations of information as communication, CAIS has provided a forum for presentation, discussion and debate.

CAIS conferences seek **submissions** related to any aspect of library and information science, particularly those which exemplify the leading edge of our discipline. **Abstracts** are refereed; **final papers** are **published in the proceedings** and **presented at the conference**. Preference is generally given to papers that report research or debate underlying methodological/philosophical issues, rather than those that report on plans yet to be implemented.

Past Conferences

2006
Information Science Revisited: Approaches to Innovation
York University, Toronto, Ontario. June 1 - 3, 2006.
Proceedings Editor: Haidar Moukdad

2005
Data, Information, and Knowledge in a Networked World
The University of Western Ontario, London, Ontario. June 2 - 4, 2005.
Proceedings Editor: Liwen Vaughan

2004
Access to Information: Technologies, Skills, and Socio-Political Context
University of Manitoba, Winnipeg, Manitoba. June 3 - 5, 2004.
Proceedings Editors: Heidi Julien and Sharon Thompson

2003
Bridging the Digital Divide: Equalizing Access to Information and Communication Technologies
Dalhousie University, Halifax, Nova Scotia. May 30 - June 1, 2003.
Proceedings Editors: Wilhelm C. Peekhaus and Louise F. Spiteri

2002
Advancing Knowledge: Expanding Horizons for Information Science
University of Toronto, Toronto, Canada. May 30 - June 1, 2002.
Proceedings Editors: Lynne C. Howarth, Christopher Cronin, and Anna T. Slawek

2001
Beyond The Web: Technologies, Knowledge and People
Université Laval, Québec, Canada. May 27 - 29, 2001.
Proceedings Editor: D. Grant Campbell

2000
Dimensions of a Global Information Science
University of Alberta, Edmonton, Alberta. May 28 - 30, 2000.
Proceedings Editor: Angela Kublik

1999
Information Science: Where Has It Been, Where Is It Going?
Université de Sherbrooke, Sherbrooke, Québec. June 9 - 11, 1999.

1998

Information Science at the Dawn of the Millennium
Université d'Ottawa, Ottawa, Ontario

1997

Communication and Information in Context: Society, Technology, and the Professions
Memorial University of Newfoundland, St. John's, Newfoundland. June 8 - 10, 1997.

1996

The Impact Of Electronic Publishing
University of Toronto, Toronto, Ontario. June 2 - 3, 1996.

1995

Connectedness: Information, Systems, People, Organizations
University of Alberta, Edmonton, Alberta. June 7 - 10, 1995.
Proceedings Editor: Hope Olson

1994

The Information Industry in Transition
McGill University, Montréal, Québec. May 25 - 27, 1994.

1993

Information As a Global Commodity: Communication, Processing and Use
St. Francis Xavier University, Antigonish, Nova Scotia. July 12 - 14, 1993.

Copies of proceedings published in print may be purchased for $40. Copies containing only abstracts are available for $10.

Proceedings volumes available for sale: 1973, 1976, 1977, 1978, 1979, 1980 (Abstracts only), 1982, 1984, 1985, 1986 (Abstracts only), 1987 (Abstracts only), 1989 (Abstracts only), 1993, 1994, 1995, 1996, 1997, 1999, 2001, 2002, 2003.

CAIS ACSI

Canadian Association for Information Science
L'Association canadienne des sciences de l'information

Home
About CAIS
CAIS Executive
Membership
Journal
Conferences
Contact CAIS
Links
Site map

Français

Conference Proceedings

Search

Last name: Year Search

Adams, Suellen S. & Kate Peirce. (2006). Is There a Transgender Canon?: Information Seeking and Use in the Transgender Community
Because transgender issues are not often openly discussed, little is known about the transgender community and its information needs. Yet, people who are dealing with transgender issues may have pressing needs. In ongoing focus group research we are exploring the ways in which members of the community have sought to meet their own information needs.
PDF Full Text

Arsenault, Clément & Élaine Ménard. (2006). Title Searches with Initial Articles: A Search Behaviour Analysis
This study examines user behaviour during know-item retrieval using title index in library catalogues. Our observations concentrate on the problems caused by the presence of an initial article or of a word homograph to an article. Measures of success and effectiveness are taken to determine if retrieval is affected in such cases.
PDF Full Text (in French)

Bartlett, Joan C. & Tomasz Neugebauer. (2006). Supporting Information Tasks with User-Centred System Design: The development of an interface supporting bioinformatics analysis
We present an interface to support the integration of bioinformatics analysis with scientific practice. The interface guides scientists through the co-ordinated use of a wide range of analyses and resources in order to solve a complex information task.
PDF Full Text

Beheshti, Jamshid, Andrew Large, Kevin Kee, & Charles Cole. (2006). Designing Virtual Environments in an Educational Context
Virtual environments in which users can navigate freely through spatial representations, pick up and examine objects, and "converse" with virtual characters, can play a role in transferring information and knowledge for both training and education. This paper discusses design issues encountered when creating such an environment for grade-five primary school students.
PDF Full Text

Boutin, Eric, Gabriel Gallezot, & Luc Quoniam. (2006). Detection of Innovation on the Web with Non-Boolean Techniques: Method, Tools and Application
The problem of the identification of weak signals is central in competitive Intelligence. The Web seems to be a very appropriate information source to detect such signals. The difficulty is the collection and the treatment of a mass and little structured data. This communication presents a hybrid non-Boolean approach that reveals innovating knowledge. The approach used information from the Internet and information collected through a thesaurus.
PDF Full Text (in French)

Bui, Yen & Jung-ran Park. (2006). An Assessment of Metadata Quality: A Case Study of the National Science Digital Library Metadata Repository
The goal of this study is to assess the quality of current metadata records in the NSDL repository. For this, we harvested over one million Dublin Core metadata records submitted through November 2005 to the repository using the Open Archives Initiative Protocol (OAIP). This study reports on the preliminary results of the tabulations and assessment of metadata quality.
PDF Full Text

Burdett, Samantha, Flis Henwood, Roma Harris, & Audrey Marshall. (2006). Mediating the 'Digital Health Divide': A Role for Public Libraries?
Preliminary findings are reported on a study of health information seeking in public libraries. The study's objectives are to identify how public libraries fit into people's information landscapes, identify challenges in supporting users in health information seeking, and contribute to the debate about the role public libraries can play in promoting healthy communities.
PDF Full Text

de Jong, Cees-Jan. (2006). Undergraduate Students' Perspectives on the Reference Transaction: A Pilot Study
A qualitative user study examining the reference transaction role in four undergraduate students'

information seeking process. Data was obtained through interviews, which were analyzed using a grounded theory approach. Findings show that participants sought no process intervention from reference librarians, had negative perceptions of the reference transaction, and valued independence during the research process.
PDF Full Text

Gazo, Dominique. (2006). The Missions of Autonomous Public Libraries in Quebec as Perceived by Elected Officials
Our doctoral research project relates to perceptions which elected officials in Quebec have about the missions of the autonomous public libraries which they are directly responsible. The emphasis is laid on the components of these missions, their significance, their legitimacy and the ideology which underlies perceptions of the elected officials.
PDF Full Text (in French)

Hayter, Susan. (2006). Exploring Information Worlds in a Disadvantaged Community: A UK Perspective
This study explored the everyday information behaviour (IB) of 21 people living on a disadvantaged housing project in the UK. The findings suggest that affective aspects of IB are key, that trust is vital and that the term 'information' can be a semantic barrier. The study further proposes ways to enable information access.

Howard, Vivian. (2006). Teens and Pleasure Reading: A critical assessment from Nova Scotia
This paper reports on the first phase of a two-part research study into the role of pleasure reading in the lives of Nova Scotia teenagers. Phase one, a quantitative survey, provides essential background context for the second phase, which uses qualitative methodology to illuminate and enrich the findings from the preliminary survey research.
PDF Full Text

Howarth, Lynne C. & Thea Miller. (2006). Assessing Metadata Categories and Visual Displays for Retrieving Digital Cultural Resources
Focus groups tested the appropriateness of a seventeen-element categorization model for uniquely identifying and retrieving digital objects from cultural repositories. Findings suggest that, while only a subset of categories ranked as important to selecting images, the type of material and a context for searching also influence the utility of a category.
PDF Full Text

Julien, Heidi. (2006). The Long Road Ahead: Information Literacy Instruction in Canada's Public Libraries
This paper reports a study of information literacy practices in Canadian public libraries. The project explored the actual and potential role of public libraries in developing the public's information literacy skills, and included a national survey of instruction and visits to public libraries where staff and library customers were interviewed.

Kipp, Margaret E. I. (2006). Complementary or Discrete Contexts in Online Indexing: A Comparison of User, Creator and Intermediary Keywords
This paper examines the context of online indexing from the viewpoint of three different groups: users, authors, and intermediaries. User, author and intermediary keywords were collected from journal articles tagged on citeulike and analysed. Descriptive statistics and thesaural term comparison shows that there are important differences in the context of keywords from the three groups.

Koshman, Sherry, Amanda Spink, Bernard J. Jansen, Chris Blakely, & Jonathan Weber. (2006). Metasearch Result Visualization: An Exploratory Study
The Missing Pieces tool visualizes the overlap of search engine results including those generated by the metasearch engine, Dogpile. The major research question is: how well can users interact with and interpret the circular metasearch results display? This study has interesting implications for the use of visualization for presenting information in a web metasearch environment.
PDF Full Text

Lambert, Frank. (2006). The mycommunityinfo.ca Approach to Online Networked Community Information Provision
Mycommunityinfo.ca offers online community information (CI) by rejecting the traditional CI directory model of excessive metadata. This dynamic and sustainable approach to providing online CI through single window access to local community and three levels of government information sources also offers an unprecedented glimpse into CI needs in Southwestern Ontario.
PDF Full Text

Li, Ping & Jamshid Beheshti. (2006). Factors Affecting Users' Mental Models of a Web Search Engine: A Case Study
Focusing on doctoral students as a specific user group, for a case study, this research investigates factors that might affect users' mental models of a Web search engine measured in the dimension of completeness, and subsequently on their search performance. Data collection techniques include interview, observation and four standard tests.

McKechnie, Lynne (E.F.), Christopher M. Dixon, Jana Fear, & Angela Pollak. (2006). Rules of (Mis)Conduct: User Behaviour in Public Libraries
Unobtrusive observation in nine sites in two public libraries in Southern Ontario explored user compliance with posted rules of conduct. With the exception of children who were frequently loud

and rambunctious, most users followed the rules. This finding is consistent with Goffman's idea that behaviour in public places is governed by normative assumptions of public order.
PDF Full Text

McKenzie, Pamela J. & Tami Oliphant. (2006). The Presentation of Complementary and Alternative Medicine Information in Canadian Midwifery Care
This paper uses discourse analysis to consider midwives' and pregnant women's discussions of conventional and complementary and alternative medicine interventions for inducing labour. Participants distinguished between "natural" and "medical" methods and used information sources based on both biomedical evidence and women's experience to justify and challenge authority claims
PDF Full Text

Ménard, Élaine, Lyne Da Sylva, & James M Turner. (2006). PeriCulture2: Using Ancillary Text for Automatic Indexing of Multimedia Objects
This paper presents the results of the project PériCulture2. The main goal of this project is to study indexing methods for Web-based non-textual cultural content. The results give an idea of the quality of the automatic indexing obtained using the ancillary text associated with multimedia objects, specifically video and sound.
PDF Full Text (in French)

Morrissey, Renée & Lisa M. Given. (2006). International Students and the Academic Library: A Case Study
This paper presents a pilot study that examined the experiences of Chinese graduate students in using the University of Alberta Libraries. The findings outline the challenges faced by students with respect to working in a second language and navigating library technologies, with a focus on the students' information literacy skills.

Nelson, Michael J. (2006). An Alternate Method for Ranking Journals Using Citations
An alternate method for ranking journals based on the algorithm used in the Google search engine for pagerank is applied to the information science and library science set of journals from Journal Citation Reports. A method of calculating individual paper influence based on this algorithm is proposed.
PDF Full Text

Nesset, Valerie. (2006). How Many Hurdles Do I Have to Jump? Conducting Research in an Elementary School Classroom
This paper examines some of the issues that may arise when conducting research in an elementary school classroom. Using examples from a recent study on the information-seeking behaviour of grade-three students, the paper identifies and discusses potential problems and barriers to the research and suggests some ways to overcome them.
PDF Full Text

Pecoskie, Jennifer L. (2006). Making Sense of the Self: Developing an Understanding of the Role of Objects within the Reading Experience
In understanding pleasure reading in everyday life we often focus on the text as a primary tool of reading and other tools, such as books, which are instrumental objects, are overlooked. This report of qualitative research focuses on the book as a tool within reading experiences and how it furthers understandings of the self for the reader.
PDF Full Text

Peekhaus, Wilhelm. (2006). Personal Health Information in Canada: Clearing the Conceptual Underbrush and Accounting for Public Opinion
This paper explores the relationship between individuals and their medical information in Canada. It employs Neill's theoretical model of privacy to situate Canadian legislation, and also analyzes public opinion data about attitudes toward medical and genetic privacy, indicating areas where legislation and public opinion are out of synch.
PDF Full Text

Qayyum, M. Asim. (2006). Improving Digital Library Interfaces by Investigating the Electronic Activities of Users
The purpose of this study was to examine the navigational patterns and text markings of electronic text readers when they interacted with electronic documents during an active reading process. The readings took place in two settings, private and document sharing, and the results provided us with user-navigational patterns taxonomy.

Rothbauer, Paulette M. & Rachelle Gooden. (2006). Representations of Young People in Information Science: The Case of the Journal of the American Society for Information Science (and Technology), 1985-2005
Using 35 articles published in JASIST between 1985 and 2005, we present the first level of our analysis of the themes we found in respect to the representation of young people as subjects of information science research. It is our general finding that the developmental approach to childhood remains dominant.
PDF Full Text

Schlemmer, Balázs. (2006). Science in the European Union – "Before and After" A Scientometric Approach to Measure the Structural Transformation of

Science in Central and Eastern Europe after the Political-Economical Changes
This scientometric study compares the scientific structures of the former EU15 and the 'newcomer' EU countries. Bibliometric indicators are used to plot the EU countries' scientific patterns based on subject fields and European co-publication maps over time. The study also investigates some peculiarities of certain EU countries' scientific journal usage.

Shiri, Ali. (2006). Knowledge Organization Systems in Canadian Digital Library Collections
The paper reports on a study of the ways in which Canadian digital library collections make use of knowledge organization systems to support users' information search behaviour. The study identified 33 digital collections which have employed some type of knowledge organization system in their search interfaces.
PDF Full Text

Singh, Rajesh. (2006). Market Orientation and Service Performance in Libraries: An Unexplored Relationship
This study investigates the inter-relationship between market orientation and service performance of 33 libraries in the south of Finland. Three kinds of market orientation were found: the strong, the medium and the weak. The findings show that the higher market orientation is positively connected with the libraries' superior service performance.
PDF Full Text

Smiraglia, Richard P. (2006). Music Information Retrieval: An Example of Bates' Substrate?
Music Information Retrieval (MIR), and ISMIR annual conferences offer a rich panoply of intellectual and cultural diversity. We map the evolution of MIR using conference papers from 2000 through 2005. Results indicate tight thematic coherence in the domain around the problems of information retrieval and classification, and the locus of most research within computer science departments.
PDF Full Text

Tennis, Joseph T. (2006). Comparative Functional Analysis of Boundary Infrastructures, Library Classification, and Social Tagging
This paper outlines three information organization frameworks: library classification, social tagging, and boundary infrastructures. It then outlines functionality of these frameworks. The paper takes a neo-pragmatic approach. The paper finds that these frameworks are complementary, and by understanding the differences and similarities that obtain between them, researchers and developers can begin to craft a vocabulary of evaluation.
PDF Full Text

Vaughan, Liwen, Margaret E.I. Kipp, & Yijun Gao. (2006). Why are Websites Co-linked? The Case of Canadian Universities
A random sample Web pages that linked to a pair of Canadian universities was retrieved. The content of the page as well as the context of the link were manually examined to record the following variables: language, country, type of Website, and the reason for co-linking.
PDF Full Text

Veinot, Tiffany, Roma Harris, Leslie Bella, Irving Rootman, & Judith Krajnak. (2006). HIV/AIDS Information Exchange in Rural Communities: Preliminary Findings from a Three Province Study
People with HIV/AIDS (PHAs) face particular challenges if they live in rural Canada, including the invisibility, stigma and limited local services. This study examines the information-seeking of PHAs and their friends/family in this rural context using three theoretical frameworks that span information seeking, incidental information acquisition and information sharing.

Yi, Kwan & Jamshid Beheshti. (2006). Boosting for Text Classification with Subject Headings
The aim of this study is to investigate how Medical Subject Headings (MeSH) as background knowledge source can improve text classification results. The hypothesis is experimented with two different sets of medical documents using HMM-based TC classifier. Experimental results show the improvement of the performance with MeSH in accuracy.
PDF Full Text

Eric Boutin, Gabriel Gallezot, Luc Quoniam
Laboratoire I3M, Université du Sud Toulon Var et Université de Nice , France

Détecter l'innovant sur le web par des techniques non booléennes : méthode, outils, application

Résumé : La problématique de l'identification de signaux faibles est centrale en Intelligence économique. Le web, par son caractère informel, apparaît comme une source d'information à privilégier. Le problème qui se pose alors est celui de la collecte et du traitement d'une information massive peu structurée. Cette communication présente une approche non booléenne hybride de révélation de connaissances innovantes combinant une recherche informelle sur le web et une recherche très structurée dans un thésaurus.

Abstract: The problem of the identification of weak signals is central in competitive Intelligence. The Web seems to be a very appropriate information source to detect such signals. The difficulty is the collection and the treatment of a mass and little structured data. This communication presents a hybrid non-Boolean approach that reveals innovating knowledge. The approach used information coming from the internet and information collected through a thesaurus.

Dans le domaine de l'intelligence économique, la détection de signaux faibles dans l'environnement d'une organisation est essentielle. La mise en évidence d'informations émergentes, le plus en amont possible, donne à l'organisation la capacité d'agir sur son environnement. Plus le signal est détecté tardivement, plus l'organisation devra subir des contraintes exogènes.
Cette communication se focalise sur l'identification de phénomènes qui ne sont pas encore émergents mais latents et qu'il convient de révéler. Il s'agit donc de « découverte de connaissances » (Knowledge Discovery in Databases)
Un certain nombre de travaux ont été proposé pour révéler des connexions à l'état latent. Ces méthodes reposent sur trois caractéristiques principales qui ont trait à la source d'information, au traitement appliqué à cette information et au domaine d'application privilégié :
- elles utilisent presque toujours une source d'information issue de bases de données bibliographiques.
- elles reposent sur une application de la logique transitive et l'exploitation de techniques de recherche non booléennes.
- elles sont surtout utilisées dans le domaine biomédical (Swanson, 1986) et peuvent, par exemple, être mises au service de la découverte de médicaments pour traiter une maladie donnée.

Dans ce travail, nous souhaitons nous abstraire de deux de œs contraintes et montrer que les méthodes qui exploitent des techniques de recherches non booléennes peuvent être utilisées :

- dans des contextes différents de celui du domaine médical
- en privilégiant non plus une information validée et structurée issue de bases de données bibliographiques mais une information hétérogène provenant d'un thésaurus et du web.

Si la détection de phénomènes latents pose un certain nombre de problèmes techniques, il s'agit avant tout d'une question de perception de l'environnement par l'homme. La méthode que nous proposons consiste en un outil de génération des possibles qui, validée ou non par

l'expert, laisse non seulement apparaître des phénomènes émergents mais peut aussi susciter chez l'expert « d'autres lumières ».
Cette communication sera organisée en 3 parties :

- Dans un premier temps, il s'agira de présenter un état de l'art des travaux réalisés dans le domaine de la découverte de connaissance dans le domaine biomédical.
- Une fois cet état de l'art effectué, nous présenterons notre méthode qui combine l'exploration d'un thésaurus et l'utilisation de classificateurs sur internet.
- Enfin, nous proposerons une illustration de la démarche autour d'une validation expérimentale : il s'agira d'identifier des indicateurs de pertinence innovants dans le domaine des moteurs de recherche.

1. Etat de l'art : découverte de connaissance dans le domaine biomédical

1.1. Découverte de connaissance dans le domaine médical : un domaine à fort potentiel

La question de la découverte de connaissances a fait l'objet, dans le domaine médical, d'un grand nombre de travaux. Il s'agit bien souvent de proposer des médicaments existant déjà sur le marché et associé à un traitement thérapeutique donné pour les faire répondre à de nouveaux traitements. Cette importance dans le domaine médical tient au potentiel économique fort de la méthode dans ce domaine. Le processus de mise sur le marché d'un médicament est un processus réglementé qui doit respecter plusieurs étapes illustrées figure 1.

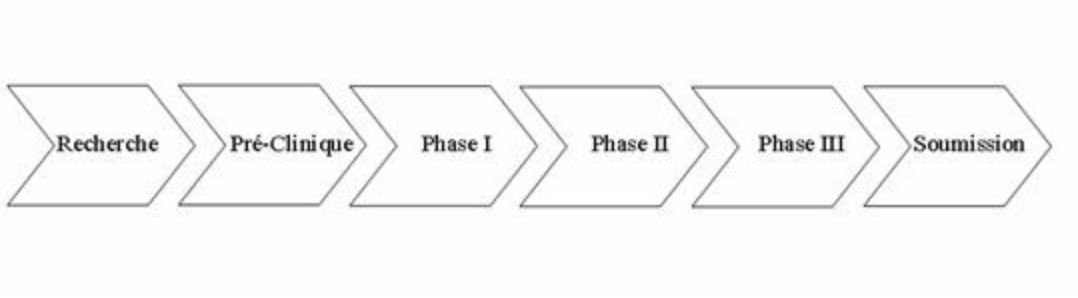

Figure 1 : Phase de recherche et développement d'un médicament

L'étape de Recherche consiste à identifier une molécule active sur une pathologie donnée
Pré-clinique : il s'agit de réaliser la formulation du principe actif et d'effectuer des études toxicologiques,
Phase I : il s'agit de montrer, chez le volontaire sain, que la molécule est bien tolérée à la dose à laquelle elle est active.
Phase 2 :l'efficacité de la molécule est testée sur de petits nombres de patients
Phase 3 : Il s'agit de confirmer à large échelle les résultats des phases II.
Soumission du dossier d'enregistrement auprès des autorités réglementaires des pays

Ce processus est long et coûteux : il est estimé à 10 ans pour un coût de 802 millions de dollars US (Lawrence, 2002), (DiMasi, 2003). Si on utilise un médicament déjà existant pour couvrir une maladie nouvelle, on gagne un temps précieux en court-circuitant le processus.

1.2. Découverte de connaissance à partir de sources de données bibliographiques : une contradiction apparente

Dans le domaine biomédical, le processus de découverte de connaissances repose sur la collecte et le traitement d'une information issue de bases de données bibliographiques. Cela constitue un paradoxe en ce sens que la finalité (l'innovation) semble en contradiction avec la source d'information privilégiée. Les bases de données bibliographiques biomédicales ne contiennent pas d'informations innovantes. Pour être référencée dans une base de données bibliographique, une publication scientifique doit suivre un processus sélectif long au terme duquel seuls quelques papiers restent en compétition.
Pour résoudre cette contradiction apparente, il faut se pencher sur la spécificité de méthode utilisée dans ces approches de découvertes de connaissances. Elles reposent en effet toutes sur des logiques non booléennes qui permettent de générer des connexions latentes insoupçonnées.

1.3. Découverte de connaissance dans le domaine médical : hypothèses sous jacentes

Pour comprendre le processus de génération d'innovation, il faut reprendre une des hypothèses communes à tous ces modèles. Tous ces modèles partent d'un constat de compartimentation de la connaissance. Le mythe du savant homme du siècle des lumières a disparu pour laisser la place à un cloisonnement des spécialités. La figure 2 empruntée à Swanson (Swanson, 1986) illustre bien ce cloisonnement des disciplines inhérent au processus de développement de connaissances nouvelles.

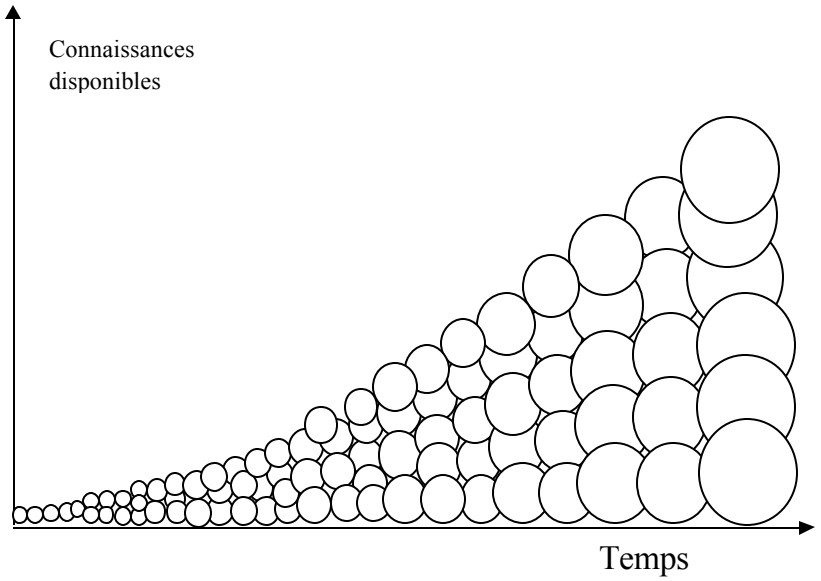

Figure 2 : La compartimentation des savoirs

Partant de ce constat, tous les modèles de découvertes de connaissance se proposent de rechercher des connexions latentes entre des disciplines devenues étanches. L'innovation ne consiste donc plus à découvrir quelque chose qui n'existait pas mais à transposer dans un autre domaine de la connaissance un phénomène déjà validé par ailleurs.

1.4. Découverte de connaissance dans le domaine médical : méthode employée.

Comment mettre en œuvre un processus de création de liens entre des disciplines étanches ? La méthode de découverte de connaissance est de ce point de vue originale par son approche et en rupture avec le modèle académique utilisé en recherche d'information. Lorsqu'on effectue une recherche d'information dans une base de données ou sur internet, on construit une équation logique simple ou élaborée (en utilisant des opérateurs booléens) qui est adressée à l'outil de recherche. Par construction, l'outil va renvoyer des documents comportant les mots de la requête. On ne trouve, par ce processus, que ce que l'on a cherché. Le mécanisme de découverte de connaissances repose lui sur l'exploitation d'une logique non booléenne.

Cette logique est basée sur la propriété de la transitivité. Dans le domaine médical, il est possible d'identifier (Swanson, 1986, 1988, 1990) 3 dimensions représentées figure 3 qui se prêtent à ce jeu transitif :

- La dimension de la maladie
- La dimension des effets physiologiques de la maladie
- La dimension des médicaments pour une maladie donnée.

Ce découpage va servir d'effet de levier pour la découverte de connaissances.

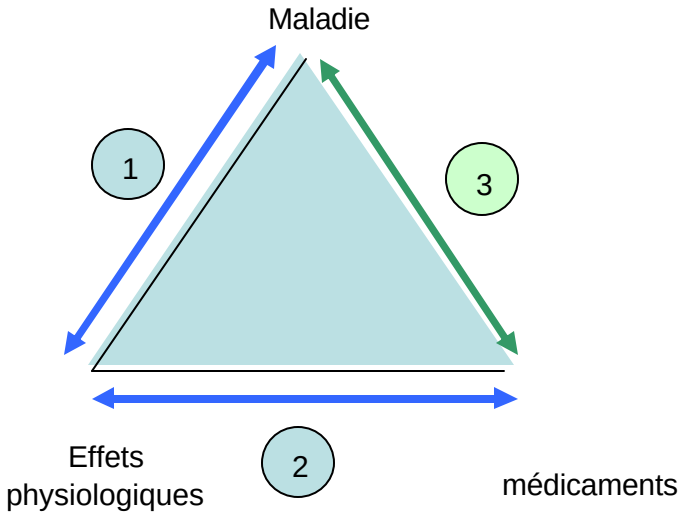

Figure 3 : principes de la logique transitive

Le mécanisme est le suivant :

- étape 1 : il est possible de connaître les effets physiologiques d'une maladie donnée. Cette opération peut être effectuée en récupérant d'une base de données biomédicale un corpus associé à une maladie donnée et à rechercher dans ce corpus quels sont les effets physiologiques associés à cette maladie.
- étape 2: Il est possible de connaître pour un effet physiologique donné le nom des médicaments actifs pour soigner cette maladie. Pour ce faire, on effectue une requête portant sur l'effet physiologique étudié dans une base de données biomédicale et on retient les médicaments les plus occurrents.
- étape 3 : Si une maladie se caractérise par un effet physiologique et si cet effet physiologique est associé à un traitement, alors, par transitivité, il est possible

d'émettre une hypothèse selon laquelle le médicament peut apparaître comme une substance permettant de gérer la maladie en question.

- Dans le plus grand nombre de cas, la relation transitive présumée est confirmée par la relation directe, le médicament étant déjà connu pour lutter contre la maladie. Dans d'autre cas, ce médicament n'est pas utilisé. Il s'agit alors d'une innovation potentielle qui doit être soumise à l'expert.

Le processus est dans la réalité plus complexe dans la mesure où une maladie se caractérise par plusieurs effets physiologiques combinés dont il faut tenir compte simultanément pour trouver le médicament adapté. Toutefois, nous souhaitons en rester ici au principe transitif simple.

Ce mécanisme transitif a été largement abordé dans la littérature biomédicale. Swanson (Swanson, 1986), le premier, a montré qu'un tel mécanisme transitif permettait de montrer que l'huile de poisson pouvait apparaître comme une substance active pour lutter contre la maladie de Raynaud. L'expérience de la maladie de Raynaud a fait l'objet de multiples réplications en utilisant des bases de données différentes, des champs différents. (Srinivasan, 2004), (Gordon, 1996), (Weeber, 2000), (Pierret, Boutin, 2004), (Smalheiser, 1998).

Dans ce travail, nous avons souhaité transposer la démarche de découvertes de connaissance au domaine du web en considérant non plus une source d'information issue de bases de données bibliographiques mais une information issue du web.

2. Transposition de la découverte de connaissances au monde du web : le modèle

2.1. Etat de l'art :

Gordon et Lindsay, qui avaient simulé les expériences de Swanson ont travaillé sur l'utilisation d'Internet dans un système de découverte de connaissances [Gordon, 2002]. Ils proposent de généraliser le modèle utilisé dans le domaine biomédical en un modèle ouvert pour la découverte de connaissances au sens large.

Gordon et Lindsay ont illustré leur démarche autour de la problématique des algorithmes génétiques. L'objectif de leur expérimentation est de trouver de nouvelles applications aux algorithmes génétiques. La méthodologie suivante est utilisée :

- les auteurs interrogent AltaVista avec la requête « genetic algorithms » et récupèrent le contenu des 50 documents les plus importants. Les termes composés de deux mots (bigrams) sont isolés et leurs fréquences calculées afin d'établir les statistiques lexicales. Douze termes en relation avec les algorithmes génétiques sont sélectionnés.
- Pour chacun de ces douze termes, une requête est adressée à Altavista. Les 100 premières réponses du moteur sont récupérées.
- Cette démarche permet de faire ressortir 42 bigrams d'une liste de 8.000, dont chacun est une découverte potentielle d'une nouvelle application des algorithmes génétiques. Par exemple, Gordon et Lindsay proposent qu'un algorithme génétique soit employé dans un modèle financier de simulation de portfolio optimisé en terme de risque et retour sur investissement.

2.2. Présentation du modèle Context – Problem - Solution:

La logique transitive repose sur la navigation entre plusieurs dimensions en s'articulant à partir d'une ou de deux dimensions pivot. Pour pouvoir généraliser la méthode et la

transposer au domaine non médical, il faut trouver trois dimensions suffisamment génériques pour ne pas se cantonner à un domaine de la connaissance particulier. Pour ce faire, nous avons conçu le modèle CPS :

- S désigne la Solution : S selon le cas correspond à la dimension outil, à l'algorithme à mobiliser, à la solution proposée pour résoudre le problème. Pour poursuivre l'exemple de Gordon et Lindsay, S désigne les algorithmes génétiques.
- P pour problème : P a une dimension fonctionnelle. Dans le contexte étudié, quel problème veut on résoudre ? Les algorithmes génétiques sont utilisés par exemple pour résoudre des problèmes d'optimisation.
- C pour contexte : C désigne la dimension applicative et correspond à un domaine de la connaissance. Les algorithmes génétiques sont utilisés par exemple dans le domaine de l'optimisation de fonctions en mathématiques.

La problématique Contexte – Problème - Solution peut être représentée par un triangle, chaque sommet illustrant une de ces dimensions.

Partant de ce triangle, notre objectif est de se servir de la dimension problème comme d'un pivot et de rechercher d'autres contextes dans lesquels la solution pourrait être utilisée ou d'autres solutions qui pourraient être employées dans notre contexte. La méthode présentée illustrée figure 4 peut donc s'appliquer à deux situations :

- trouver une solution nouvelle à un problème existant. A partir d'un triangle « est » qui correspond au point de départ de l'expert, l'idée est de s'ouvrir à d'autres réalités, en l'occurrence le triangle « ouest ». Triangle « est » et « ouest » ont le sommet problème en commun. Le problème est donc le même mais la solution et le contexte sont différents. Peut être la solution B apparaîtra-t-elle comme pertinente pour le contexte A recréant ainsi un nouveau triangle dont l'arc manquant est représenté au nord en pointillé. Ce genre de navigation n'est pas simple car les problèmes ne sont pas forcément formulés de la même manière avec un même vocabulaire dans différentes disciplines : cette méthode suppose un gros travail de transcription pour que des problèmes, apparaissant comme identiques, se recouvrent sous la même acception.
- déterminer un contexte applicatif nouveau à une solution éprouvée. Dans ce cas, il s'agit de valoriser la solution A dans d'autres contextes. Dans le cas de la figure 4, on peut identifier un nouveau contexte (B) à travers le triangle « sud » défini avec des pointillés.

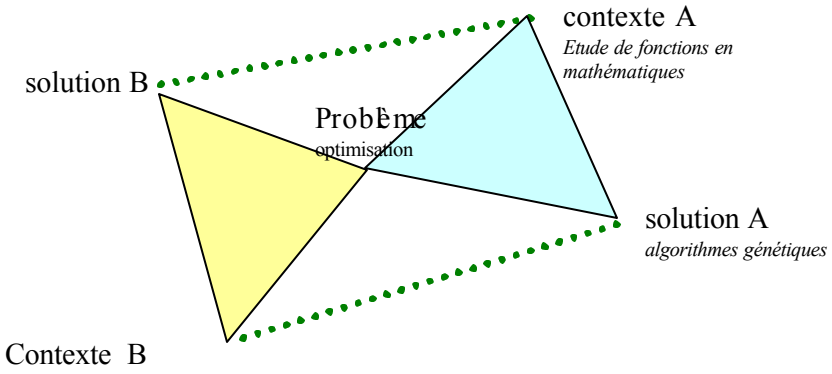

Figure 4 : le modèle CPS : transitivité potentielle des contextes et des solutions

2.3. Présentation de la démarche et des outils de traitement :

Dans le domaine médical, on dispose d'une information structurée et validée et du thésaurus Mesh. L'exploitation de ce thésaurus permet de récupérer trois listes fermées :

- Une liste de maladies
- Une liste d'effets physiologiques
- Une liste de médicaments

Ces listes sont extrêmement précieuses car elles permettent d'extraire des notices bibliographiques les informations appartenant à l'une ou l'autre de ces dimensions. Au lieu de travailler en texte intégral, on travaille sur les termes thésaurés du Mesh refermés sur une de ces trois dimensions.

Nous ne disposons pas d'un thésaurus général composé de trois listes correspondant aux trois dimensions du CPS. La question qui se pose est alors celle de la navigation entre des ensembles hétérogènes qui n'ont pas forcément de langage commun. La découverte de connaissances sur Internet se trouve confronté au problème du vocabulaire. Ceci impose d'employer des concepts ou descripteurs génériques afin de rendre leurs contenus plus facilement manipulables ou de connaître les divers termes correspondant à un problème donné.

La démarche que nous proposons s'articule autour de trois étapes alternant une démarche exploratoire formelle et informelle. La première est une étape de stimulation de la créativité de l'expert. La seconde très formelle consiste à naviguer au sein d'un thésaurus. La dernière consiste à explorer les ressources web :

- Etape 1 : « générique » le problème

La première étape consiste à partir de la Solution pour remonter au problème : Cette étape consiste à rendre générique le problème en essayant de s'abstraire du domaine d'application choisi pour voir quel problème il peut résoudre. Ce travail peut être réalisé en consultant l'expert du domaine. A l'issue de cette étape, on dispose d'une liste de mots clés. Ces mots clés décrivent souvent plusieurs facettes du phénomène à analyser. Pour suivre l'exemple précédent, on part des algorithmes génétiques et on essaie d'identifier quels problèmes ils essaient de résoudre.

- Etape 2 : exploration à partir d'un thésaurus

Cette deuxième étape prend pour point de départ le résultat de l'étape antérieure. Chaque mot clé identifié par l'expert est introduit dans le thésaurus français Rameau (Répertoire d'Autorité-Matière Encyclopédique et Alphabétique Unifié http://rameau.bnf.fr). Ce thésaurus généraliste multidisciplinaire est commun non seulement à l'ensemble des Bibliothèques Universitaires mais aussi à la Bibliothèque Nationale de France, et à d'autres bibliothèques de lecture publique. L'exploration de ce thésaurus va permettre, pour un mot clé donné de chercher les termes parents, enfants et frères de ce mot clé. On identifie ainsi la famille de chaque mot clé de l'étape précédente. Cette étape élargit donc la liste des mots clés. Le fait de considérer les mots clés parents permet de remonter à un niveau de généralisation et d'abstraction. On arrive ainsi à s'abstraire du problème courant et à le rendre générique. A l'issue de cette étape, on obtient une liste d'associations entre les mots clés à travers les relations hiérarchiques qui les lient au sein du thésaurus. Ces relations peuvent faire l'objet de représentations cartographiques.

- Etape 3 : identifier des connections latentes à partir du web

Après avoir élargi la première liste de mots clé à l'aide du thésaurus, la liste des termes thésaurés est soumise à l'expert. Celui-ci va choisir un certain nombre de termes qui peuvent

correspondre à des termes de la liste initiale complétée par des termes nouveaux issus du thésaurus utilisé.
Ces termes sont ensuite injectés dans un classificateur web de type www.grokker.com Le fait de passer par cette étape de classification automatique de corpus web présente deux avantages :

- s'abstraire du thésaurus très formalisé, permettant ainsi de générer des possibles insoupçonnés
- Le classificateur automatique accepte en entrée non plus des mots clés isolés mais des associations de mot clés, permettant d'analyser l'interaction possible entre des domaines de la compétence étanches.

3. validation expérimentale :

3.1. énoncé du problème :

Notre expérimentation porte sur les indicateurs de pertinence des moteurs de recherche. Notre question est: « Peut on découvrir un indicateur ou une famille d'indicateur de pertinence de moteur de recherche innovant à partir d'une exploration de ce qui se passe dans d'autres contextes ? ». Le problème peut être formulé en utilisant le triangle CPS présenté figure 5.

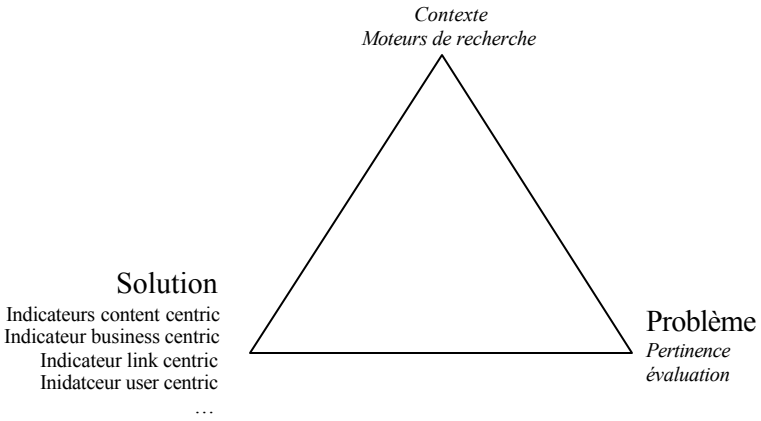

Figure 5 : triangle CPS adapté au contexte des indicateurs de pertinence

3.2. méthode de résolution

Etape 1 : remonter au problème en le rendant générique : Dans cette étape, il s'agit de trouver le dénominateur commun entre tous les indicateurs de pertinence de moteurs de recherche. Traditionnellement les algorithmes de pertinence des moteurs de recherche recouvrent différentes familles de technologies. Chacune repose sur des hypothèses sous-jacentes concernant l'autorité qui confère à une page web sa pertinence.

⇒ Indicateur de pertinence de type "Content centric" : cette famille d'indicateurs apprécie la pertinence d'une page web pour un sujet donné par l'analyse de son contenu.

⇒ Indicateur de pertinence de type "Link centric" : L'analyse relationnelle est une analyse qui va qualifier la pertinence d'une page par sa capacité à obtenir un nombre significatif de liens entrants de qualité en provenance d'autres pages du web.

⇒ Indicateur de pertinence de type "Business centric" : Cette approche "business oriented" est privilégiée par les moteurs de recherche payants ou par les zones payantes des moteurs de recherche. La pertinence d'une page dépend alors de la somme d'argent que le concepteur est prêt à payer pour que sa page soit bien référencée.

⇒ Indicateur de pertinence de type "User centric" : Cette approche a été introduite par le moteur Directhit qui considérait que c'est le temps passé par les internautes sur une page qui va servir d'indicateur de pertinence à cette page. Selon cette approche, l'internaute (le client du moteur) a un rôle fort dans la définition de la pertinence d'une page web.

Les indicateurs de pertinence présentés ci-dessus s'appliquent à des corpus composés de pages web. Ces corpus présentent différentes caractéristiques :

- Les pages web sont interconnectées
- Il y a une logique d'usage car les pages web sont parcourues par des internautes
- Il y a une logique thématique car ces pages web traitent de certains sujets
- Il y a une logique de valeur car l'information a une valeur pour celui qui la possède
- Il y a une logique de traçabilité des usages

Pour autant, toutes ces caractéristiques ne sont pas mobilisées au même titre dans chaque indicateur de pertinence. Google, dans son Pagerank privilégie l'interconnexion entre pages web, les indicateurs de type « user centric » privilégient les usages et le caractère interconnecté des pages.

On s'aperçoit aussi que chaque indicateur de pertinence revient finalement à privilégier un évaluateur qui peut être les pairs, le marché, l'usager, le site lui-même. Si on veut trouver de nouveaux indicateurs de pertinence de moteur de recherche, une piste peut consister à identifier de nouvelles *instances évaluatrices ou instances attributives de l'évaluation***.**

Principale instance attributive de la pertinence	Exemple d'indicateur de pertinence
Un expert	Annuaire de recherche
Un robot	Moteur de recherche
Les pairs	Critères exogènes : algorithme de type Pagerank de Google
Le marché	Un algorithme business oriented
Auto définition de la pertinence	Critère endogène : analyse de contenu
L'usager	Indicateur user centric type directhit
Le Hasard (et sélection] ou subjectivité ou abduction de Pierce	Serendipité (expertise ou subjectivité de l'usager)
A trouver	L'innovation que nous recherchons

Tableau 1 : les instances attributives de l'évaluation

Si on peut montrer que, dans d'autres domaines de la connaissance, il existe une instance évaluatrice d'un autre type, alors cette instance pourra être transposée au domaine du web.

Pour trouver un nouvel algorithme de pertinence de moteur de recherche, il faut revenir un cran en arrière et identifier les mots clé génériques caractéristiques de cette logique d'indicateur de pertinence. Une première liste de mots clés a été identifiée : *classement, pertinence, réseau, évaluation, algorithme, moteur de recherche***.** (ranking criteria, relevance criteria, relevance ranking algorithm, relevance indicator, network, evaluation, algorithm, search engine).

Etape 2 : le recours au thésaurus :

Nous avons sollicité le thésaurus français Rameau en lui injectant successivement les mots clés identifiés lors de l'étape précédente. Pour chaque mot clé, nous recherchons :

- ses parents
- ses enfants
- ses frères et sœurs.

On reconstitue donc pour chacun des mots clés sa famille au sens strict. On obtient donc plusieurs familles : *classement, pertinence, réseau, évaluation, algorithme, moteur de recherche*

Ces différentes familles sont alors représentées sur une cartographie représentant les relations hiérarchiques entre mots clés. Un exemple d'une telle cartographie est proposé figure 6.

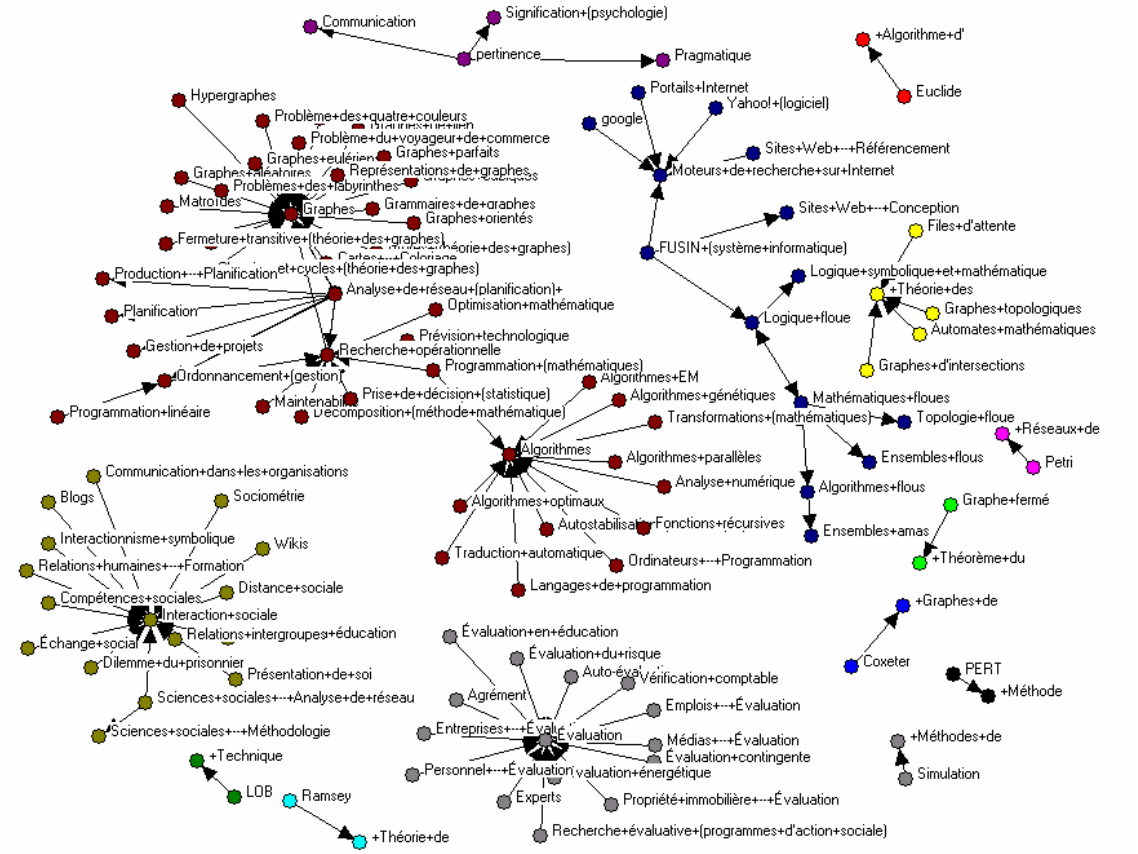

Figure 6 : visualisation des interactions à partir d'une navigation sur Rameau

Cette cartographie se présente sous forme d'un graphe composé de plusieurs composantes fortement connexes. Chaque composante fortement connexe correspond peu ou prou à un mot clé entré dans Rameau. On peut voir dans ce graphe que la partie marron au Nord ouest est composée de deux blocs normalement disjoints. Le bloc construit autour du mot clé « algorithme » se trouve connecté au bloc construit autour du mot clé « analyse de réseau (planification) » par l'intermédiaire du mot clé « recherche opérationnelle ». Cette observation peut être une piste intéressante. Les moteurs de recherche n'utilisent pas d'algorithmes issus de techniques de recherche opérationnelle. Pourtant ces techniques sont utilisées dans des problèmes d'analyse de réseau en planification. Sans doute faudrait il soumettre la problématique de la pertinence à un expert de « recherche opérationnelle ».

Ce graphe nous permet aussi de générer de nouveaux mots clés pertinents qui ne faisaient pas partie de la liste initiale : hypergraphes, sociométrie, topologie floue. Ce terme de sociométrie

est intéressant car l'indicateur que nous recherchons est un outil automatique de traitement d'une information réseau : la sociométrie dispose sans doute d'indicateurs à explorer.

Etape 3 : identifier des connections latentes à partir du web :
Nous allons partir d'une liste de termes charnières identifiés lors des deux étapes précédentes et analyser l'articulation qui peut exister entre ces mots charnières. Pour se dispenser d'une analyse textuelle trop lourde, nous avons raisonné sur un classificateur automatique : Grokker. Les mots clés que nous avons retenus appartiennent à plusieurs registres. Ils sont représentés tableau 2

network	analysis
Social network	algorithm
sociometry	criteria
hypergraph	Indicators
Link	Measures
Operational research	relevance
	ranking
	evaluation

Tableau 2 : mots clés à considérer

Plusieurs couples de mots clés ont été soumis à Grokker. Dans tous les cas, les résultats ont été analysés à la recherche de connections nouvelles.
Nous allons décrire une des pistes potentielles que nous avons identifiée. Elle a consisté à partir de la requête « sociometric network ». Grokker a suggéré « sociometric measures » . Nous avons ensuite été conduit à « centrality measures » pour être orienté vers des critères de centralité de type (degree centrality, closeness centrality, information centrality). En approfondissant la recherche et en sollicitant un expert en sociométrie, on se rend compte que la centralité de degré («degree centrality ») correspond à l'indicateur de pertinence de moteur de recherche appelé popularité (nombre de liens entrant sur une page). Cet indicateur exploité dans le domaine des réseaux sociaux a donc été transféré dans le domaine du web. Il a été abandonné car trop facilement spammable. Par contre les autres indicateurs de centralité en usage, en analyse des réseaux sociaux et en sociométrie ne sont pas exploités dans le domaine des indicateurs de pertinence des moteurs de recherche. Plusieurs documents parlent de ces indicateurs de centralité dans des contextes web plus pour décrire l'organisation des données sur le web que pour s'en servir pour déterminer le classement de telle ou telle page. Il y a potentiellement là des éléments qui pourraient faire l'objet d'une transposition au monde des indicateurs de pertinence des moteurs de recherche.

La découverte de connaissance est un domaine complexe qui nécessite des compétences d'interface et un esprit ouvert. Dans un monde complexifié, la découverte de connaissances suppose la maîtrise d'une approche qui soit capable de filtrer les données et de suggérer un petit nombre d'associations innovantes qui puissent être proposées à l'expert pour validation.
L'approche que nous proposons est une approche duale qui va se nourrir de la complémentarité de la rigueur du recours à un thésaurus et d'une exploitation de ressources web par l'intermédiaire d'un classificateur de données web.
Cette chaîne de traitement de l'information est pour l'instant semi automatique. Il nous semble délicat d'envisager une procédure automatique permettant de gérer ce processus qui intègre à chaque maillon de la chaîne l'intervention de l'expert du domaine.

Bibliographie :

DiMasi, J.A., Hansen, R.W, Grabowski, H.G. (2003), "The price of innovation: new estimates of drug development costs", Journal of Health Economics. Vol. 22, n°2, p. 151-185.

Gordon, M.D., Lindsay, R.K. (1996), "Toward discovery support systems: a replication, re-examination, and extension of Swanson's work on literature-based discovery of a connection between Raynaud's and fish oil", Journal of the American Society for Information Science. Vol. 47, n°2, p. 116-128.

Gordon, M., Lindsay, R.K, Fan, W. (2002), "Literature-based discovery on the World Wide Web", ACM Transactions on Internet Technology. Vol. 2, n°4, p. 261-275.

Lawrence, R.N. (2002), "Sir Richard Sykes contemplates the future of the pharma industry", Drug Discovery Today. Vol. 7, n°12, p. 645-648.

Pierret, J.D., Boutin E. (2004), "Découverte de connaissances dans les bases de données bibliographiques. Le travail de Don Swanson : de l'idée au modèle", ISDM, n°109.

Smalheiser, N.R., Swanson D.R. (1998), "Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses", Computer Methods and Programs in Biomedicine. Vol. 57, n°3, p. 149-153.

Srinivasan, P. (2004), "Text mining: generating hypotheses from MEDLINE", Journal of the American Society for Information Science. Vol. 55, n°5, p. 396-413

Swanson, D.R. (1986), "Fish oil, Raynaud's syndrome, and undiscovered public knowledge", Perspectives in Biology and Medicine. Vol. 30, n°1, p. 7-18.

Swanson, D.R. (1988), "Migraine and magnesium : eleven neglected connections", Perspectives in Biology and Medicine. Vol. 31, n°4, p. 526-557.

Swanson, D.R. (1990), "Somatomedin C and arginin : implicit connections between mutually- isolated literatures", Perspectives in Biology and Medicine Vol. 33, n°2, p. 157-186.

Weeber, M.A., Klein, H., Aronson, A.R., Mork, J.G., de Jong – van den Berg, L.T.W., Vos, R. (2000), "Text-based discovery in biomedicine: the architecture of the DAD-system", Proceedings of the AMIA Symposium. p. 903-907.