Thésaurus et clusterisation automatique de données web : deux outils au service de la détection de signaux faibles

Eric Boutin, Luc Quoniam

Thésaurus et clusterisation automatique de données web : deux outils au service de la détection de signaux faibles

Boutin Eric
Université du Sud Toulon Var
Maître de Conférences laboratoire I3M IUT TC
BP 132 83957 la Garde Cedex
boutin@univ-tln.fr
+33 4 94 14 23 56

Quoniam Luc
Université du Sud Toulon Var
Professeur des Universités
Institut Ingémédia
BP 132 83957 la Garde Cedex
quoniam@univ-tln.fr

Resumo :
A identificação de sinal fraco e crucial no processo de inteligência econômica. A Web, parece uma fonte a privilegiar, como fonte de informação informal. O problema vem a seguir da coleta e tratamento de uma grande quantidade de informação pouco estruturada.
Este trabalho apresente uma solução híbrida "não boolean" de revelação de conhecimento inovador. A solução testada combina uma busca de informação informal na Web e uma busca bastante estruturada com thesaurus.

Résumé :
La problématique de l'identification de signaux faibles est centrale en Intelligence économique. Le web, de part 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 et peu structurée.
Cette communication présente une approche non booléenne hybride de révélation de connaissances innovante. L'approche utilisée combine 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 an hybrid nonBoolean 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 s'agit 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 ces 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.
- Une validation expérimentale nous permettra d'illustrer la démarche et de juger de sa recevabilité.

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.

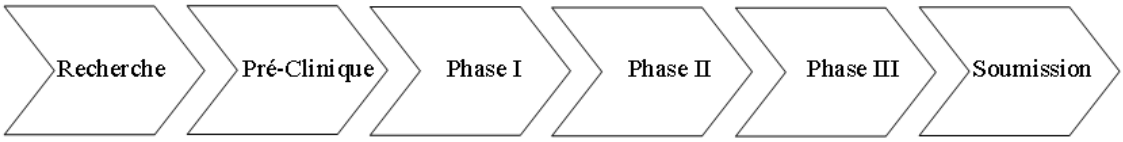

Figure 1 :Phase de recherche et développement d'un médicament

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. [Lawrence, 2002], [DiMasi, 2003] l'estiment à 10 ans pour un coût de 802 millions de dollars US. 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 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 illustre bien ce cloisonnement des disciplines inhérent au processus de développement de connaissances nouvelles.

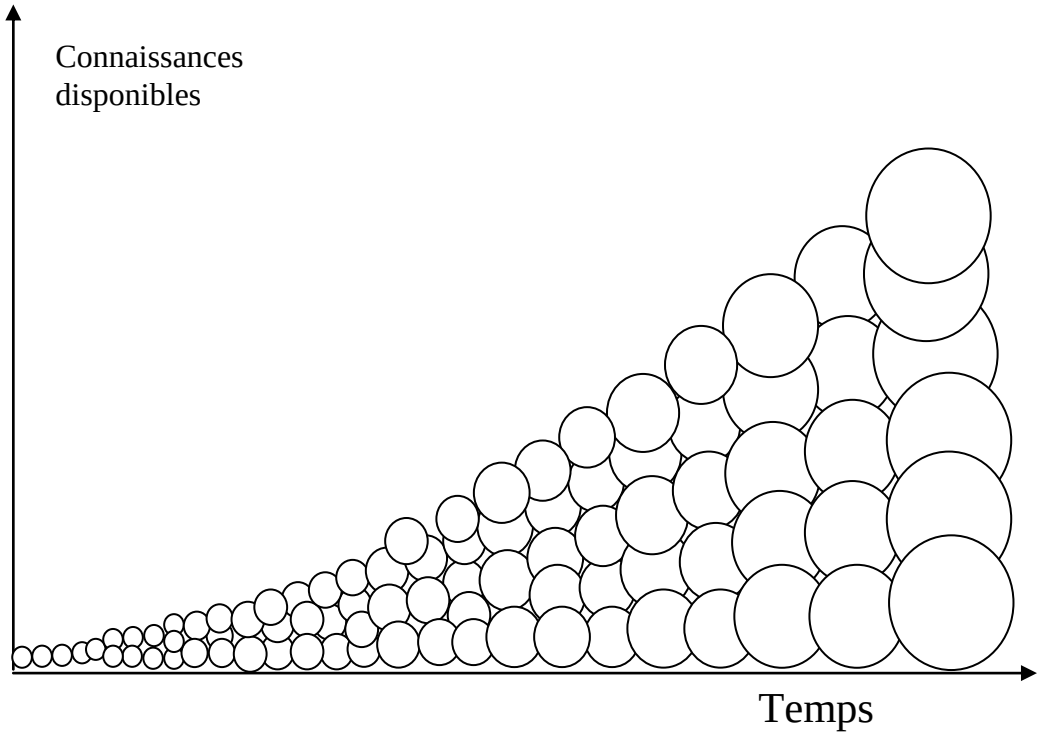

Figure 2 : La compartimentation des savoirs

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 re-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.

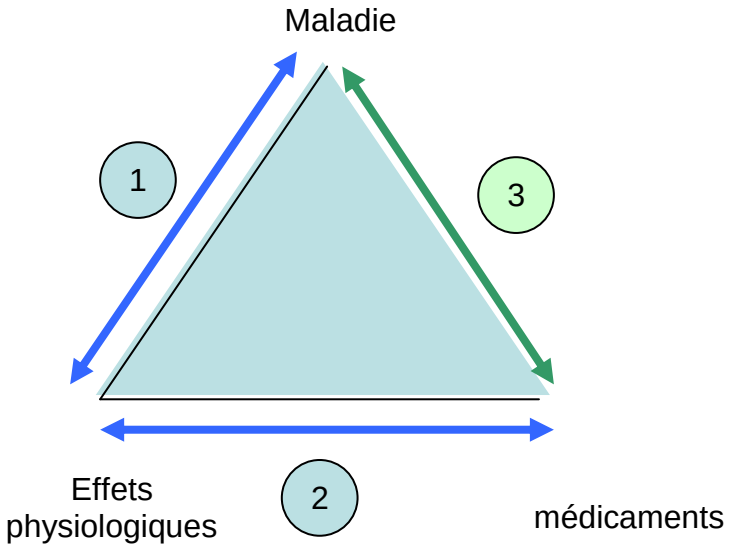

Figure 3 : principes de la logique transitive

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 (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 :

- C pour contexte : C désigne la dimension applicative et correspond à un domaine de la connaissance
- P pour problème : P a une dimension fonctionnelle. Dans le contexte étudié, quel problème veut on résoudre ?
- S pour Solution : S selon le cas correspond à la dimension outil, à l'algorithme à mobiliser, à la solution proposée pour résoudre le problème.

La problématique Contexte – Problème- Solution peut être représentée par un triangle, chaque sommet illustrant une de ces dimensions.
L'objectif de la méthode est de voir si dans d'autres contextes, le même problème se pose et de voir si d'autres solutions y sont apportées. Pour reprendre notre analogie graphique, il

s'agit de se servir de la dimension problème comme d'un pivot et de rechercher d'autres contextes dans lesquels d'autres solutions pourraient être apportées. La méthode présentée illustrée figure 4 peut s'appliquer à deux situations :

- trouver une solution nouvelle à un problème existant. A partir d'un triangle jaune qui correspond au point de départ de l'expert, l'idée est de s'ouvrir à d'autres réalités, en l'occurrence le triangle bleu. Le triangle bleu a comme point commun avec le triangle jaune le sommet « Problème ». 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 blanc. 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 2, on peut identifier un nouveau contexte (C) à travers le triangle vert. Ce problème est a priori moins difficile à résoudre que le précédant car ce triangle a une arête commune avec le précédant et non plus seulement un sommet.

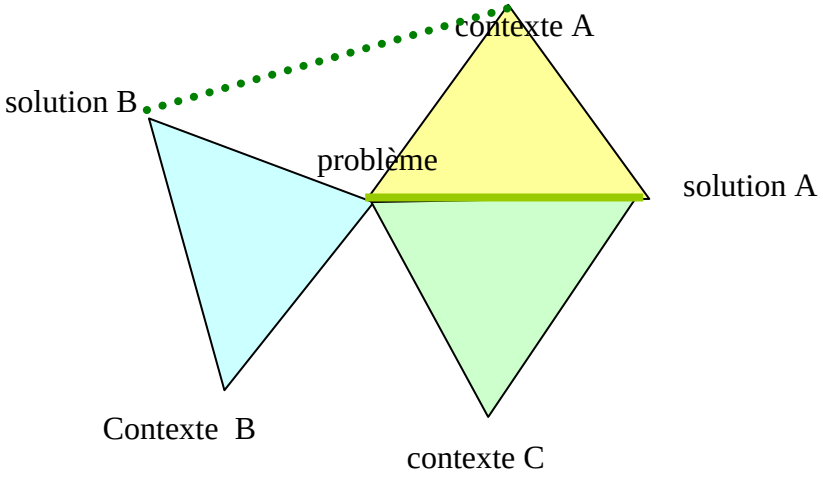

Figure 4 : le modèle Contexte – Problème- Solution

A partir d'un seul triangle de départ, il est donc possible, du moins en théorie, à l'issue d'un processus de génération de solutions nouvelles, de reconstituer une « pale de moulin à vent » dont le rotor serait le problème à résoudre.

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 du 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 des 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 : généricuer le problème

La première étape consiste à partir de la Solution pour remonter au problème : Cette étape consiste à généricuer 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.

- 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, pour chaque mot clé de l'étape précédente sa famille. 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 généricuer. 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 ? »

3.2. méthode de résolution

Etape 1 : remonter au problème en le génériquant : 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 aune 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	**serendipité**
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 clé 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 5.

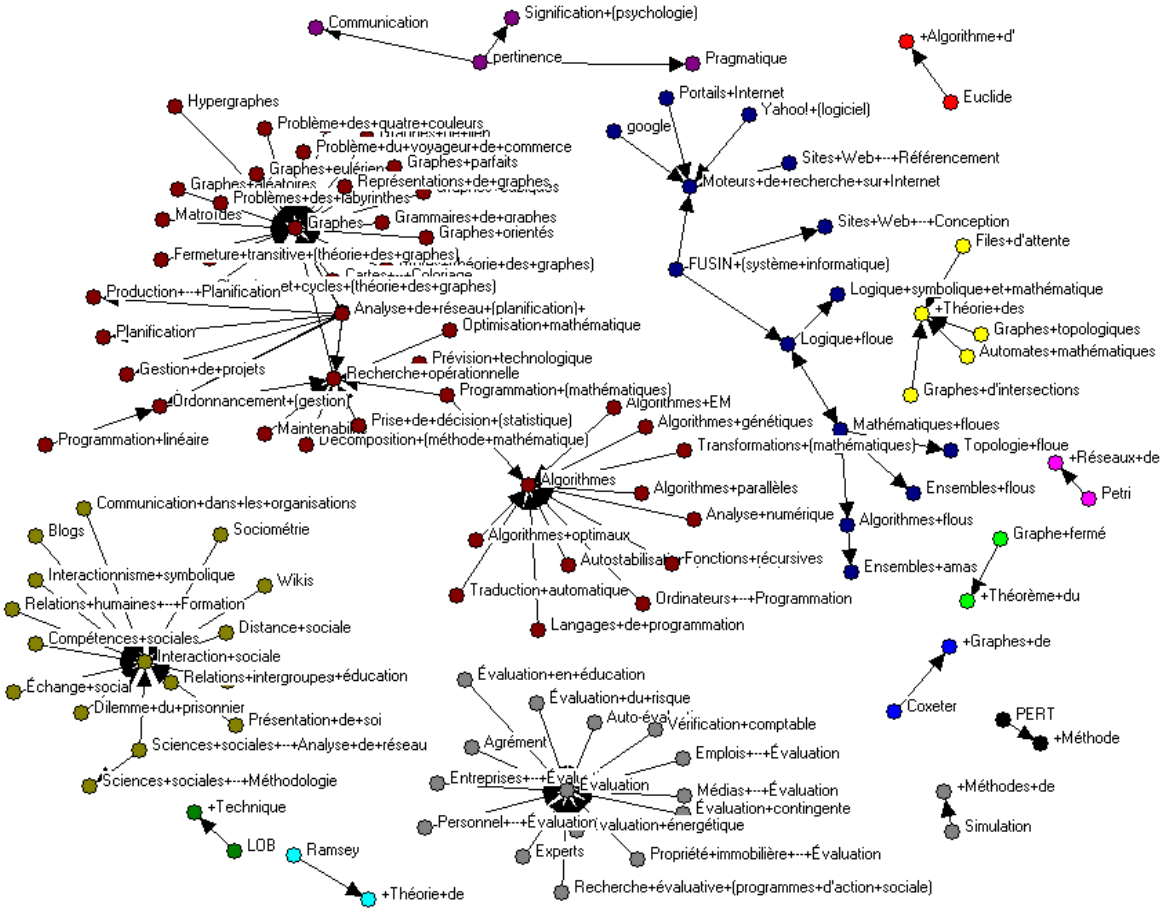

Figure 5 : visualisation des interactions à partir d'une navigation sur Rameau

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.

<u>Etape 3 : identifier des connections latentes à partir du web:</u>
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.

Conclusion :

La découverte de connaissance est un domaine complexe qui nécessite des compétences d'interface et un esprit ouvert. Dans un monde de plus en plus complexe, 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 un 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.