Terminologie & Ontologie : Théories et applications

Actes de la conférence

TOTh 2010

Annecy – 3 & 4 juin 2010

avec le soutien de :

- Ministère de la Culture et de la Communication, Délégation Générale à la Langue Française et aux Langues de France
- Association Européenne de Terminologie
- Société française de terminologie
- Ecole d'ingénieurs Polytech'Savoie – Université de Savoie
- Université de Sorbonne nouvelle
- Association EGC (Extraction et Gestion des Connaissances)
- ISKO (International Society for Knowledge Organization) France

Institut Porphyre
Savoir et Connaissance

`http://www.porphyre.org`

Comité scientifique

Président du Comité Scientifique : Christophe Roche

Comité de pilotage

Loïc Depecker	Professeur, Université de Sorbonne nouvelle
André Manificat	Directeur, GRETh
Christophe Roche	Professeur, Université de Savoie
Philippe Thoiron	Professeur émérite, Université de Lyon II

Comité de programme

Bruno de Bessé	Professeur, Université de Genève
Franco Bertaccini	Professeur, Université de Bologne
Gerhard Budin	Professeur, Université de Vienne
Marc van Campenhoudt	Professeur, Termisti, ISTI, Bruxelles
Danielle Candel	CNRS, Université Paris Diderot
Stéphane Chaudiron	Professeur, Université de Lille 3
Rute Costa	Professeur, Universidade Nova de Lisboa
Luc Damas	MCF, Université de Savoie
Sylvie Després	Professeur, Université Paris 13
François Gaudin	Professeur, Université de Rouen
Anne-Marie Gendron	Chancellerie fédérale suisse, Section terminologie
Jean-Yves Gresser	Ancien Directeur à la Banque de France
Ollivier Haemmerlé	Professeur, Université de Toulouse
Michèle Hudon	Professeur, Université de Montréal
John Humbley	Professeur, Université Paris 7
Michel Ida	Directeur MINATEC, CEA
Hendrik Kockaert	Professeur, Lessius Hogeschool (Anvers)
Michel Léonard	Professeur, Université de Genève
Pierre Lerat	Professeur honoraire, Equipe Condillac
Widad Mustafa	Professeur, Université de Lille 3
Fidelma Ní Ghallchobhair	Foras na Gaeilge (The Irish-Language Body)
Henrik Nilsson	Terminologicentrum TNC, Suède
Jean Quirion	Professeur, Université d'Ottawa
Renato Reinau	Suva, Lucerne
François Rousselot	MCF, Université de Strasbourg
Gérard Sabah	CNRS, Orsay
Michel Simonet	CNRS, Grenoble
Marcus Spies	Professeur, Université de Munich
Dardo de Vecchi	Professeur associé, Euromed-Management

Comité d'organisation :

Responsable : Luc Damas

Samia Chouder, Joëlle Pellet

Avant propos

Cette année la conférence a été précédée d'une journée de formation consacrée à la terminologie et l'ontologie, à leurs liens et leurs apports mutuels. L'intérêt qu'a suscité cette journée nous amènera certainement à réitérer l'opération les années suivantes.

Le succès de la conférence d'ouverture de notre collègue Frédéric Nef, portant sur l'ontologie prise dans sa dimension philosophique, a montré, s'il en était encore besoin, la richesse d'une approche pluridisciplinaire.

Animées par différents présidents, les sessions ont alterné présentations théoriques et démonstrations de systèmes, offrant ainsi l'opportunité à plusieurs industriels de nous parler de leurs projets. L'éventail des sujets abordés, à travers les quatorze présentations retenues (incluant la conférence d'ouverture) réparties sur deux jours, illustre la richesse mais aussi la vitalité de notre communauté : aide à la traduction, thésaurus multilingue, phraséologie, entité nommé, recherche d'information, etc. L' « actualité » n'a pas été oubliée à travers une ontologie des risques financiers.

Enfin, les Conférences TOTh sont devenues internationales à partir de cette année avec le français et l'anglais comme langues officielles. Le comité de programme s'est ouvert à de nouveaux membres portant à dix le nombre de pays représentés et à plus de 40% le nombre de personnalités étrangères. Gageons que cette ouverture sera prometteuse.

Christophe Roche

Président du Comité Scientifique

Table des matières

CONFERENCE INVITEE

ARTICLES

DEMONSTRATIONS

L'ontoterminologie pour la recherche d'information sémantique

Luc Damas, Christophe Tricot

Résumé : En dehors d'une gestion électronique de documents à base de thesaurus, les nouvelles technologies (linked data) amènent de nouveaux usages et de nouveaux utilisateurs. Les bases documentaires grandissent de pair avec les capacités de stockage. Classer les documents manuellement devient irréalisable et trouver une information se complexifie. La recherche par mots-clés s'impose comme une alternative incontournable à la navigation. Les moteurs de recherche actuels, connus principalement sur Internet, se basent sur les mots-clés pour fournir des documents répondant aux besoins de l'utilisateur. Si le besoin se fait précis, généralement dans un contexte métier, l'imprécision de la recherche devient gênante, autant par l'apparition de documents hors-sujet que par l'absence de certains documents pertinents. La gestion sémantique de ces documents devient nécessaire. L'ontoterminologie, regroupant une dimension conceptuelle et une dimension terminologique, est une théorie intéressante sur laquelle les processus documentaires peuvent s'appuyer : La compréhension des concepts du domaine et les mots pour en parler.

Mots-clés : Recherche d'information sémantique, ontoterminologie

1. Introduction

La taille des bases documentaires augmente de pair avec les capacités de stockage. Les usages ont évolué vers le tout électronique, sans souci de quantité. Tout est conservé. Le classement manuel de ces documents est une tâche qui dépend de nombreux facteurs et dont le résultat peut varier fortement. Pour une même personne, un document peut avoir de multiples utilités, et se retrouver dans des catégories différentes en fonction des besoins. De manière collaborative, au sein d'une même organisation, plusieurs personnes peuvent posséder des points de vue variés, et donc classer un même document de manières radicalement différentes. Enfin, un même document, en fonction de son volume, peut traiter de plusieurs thèmes et ainsi appartenir à plusieurs catégories. Ces considérations ont mené à une variation des usages, qui deviennent de moins en moins normatifs. Le classement des documents devient sommaire et les utilisateurs se fient d'avantage aux outils de recherche.

Cette tendance est confortée par une habitude de plus en plus grande à effectuer des requêtes sur Internet. Formuler un besoin de manière textuelle

plutôt que de naviguer dans un espace documentaire devient la norme, au sein d'un système d'information, mais aussi, de manière individuelle, au sein d'un logiciel de gestion de courriels ou d'un stockage physique sur disque dur.

Un système de recherche d'information (SRI) cherche à mettre en correspondance des besoins plus ou moins bien exprimés par un individu et des réponses à ces besoins (Chevallet, 2009). La formulation de la requête est un processus cognitif dans lequel l'individu exprime sous la forme de mots une compréhension d'un thème. Le défi consiste à lui fournir les documents correspondant à cette compréhension.

Notre approche se base sur l'ontoterminologie (Roche, 2007). L'avantage du modèle ontoterminologique est de distinguer clairement les aspects conceptuels, regroupant et structurant les idées, des aspects terminologiques, regroupant les mots et usages. Dans notre vision d'une recherche d'information (RI) qui vise à combler un « besoin cognitif » exprimé à l'aide de mots, l'ontoterminologie semble adaptée.

L'article présente dans un premier temps les éléments fondamentaux de la RI, de l'indexation à la recherche en passant par les mesures de pertinence. Nous présentons ensuite l'ontoterminologie en nous appuyant sur un exemple concret. Nous détaillons enfin notre système de recherche sémantique en l'illustrant d'exemples caractéristiques.

2. Eléments de RI

2.1. Définitions et principes généraux

Un Système de Recherche d'Information (SRI) est un outil informatique permettant la mise en correspondance d'un besoin, exprimé sous la forme d'une requête, et d'un élément d'information (*Information item* (Baeza-Yates et. al, 1999). Un élément d'information, au sens des Systèmes d'Information, est tout élément enregistré dans le système, en général un fichier : un document PDF, une page web, une image, une vidéo… La recherche d'information fournit donc des documents, conteneurs de l'information recherchée. La différence entre recherche documentaire et recherche d'information est ténue, et provient principalement des différentes disciplines qui manipulent les documents (bibliothécaires, informaticiens). Ainsi, à la requête « quel temps

fera-t-il demain ? », un SRI ne dira pas « Il fera beau », mais : « Cette information est disponible dans tel document »[1].

a) Requête

Une requête est généralement un ensemble de mots. Sous cette forme non ordonnée, il est supposé que l'utilisateur cherche un document contenant tous les mots. La plupart des moteurs de recherche proposent une syntaxe simple à base d'opérateurs logiques permettant d'affiner une recherche. La conjonction (ET logique) est la formulation d'une recherche de tous les mots. Elle est la forme par défaut (sans opérateur) proposée par les moteurs actuels. La disjonction (OU logique inclusif) est la formulation d'une recherche d'un mot parmi une liste de mots. Dans google par exemple, le symbole à utiliser est `or`. L'expression « terminologie or ontologie » permet de trouver tous les sites web contenant le mot « terminologie » ou le mot « ontologie » ou les deux. Enfin, la négation permet de trouver des résultats ne contenant pas un mot donné. L'expression « terminologie –ontologie » est une requête permettant d'accéder aux documents contenant le mot « terminologie » et ne contenant pas « ontologie ».

b) Recherche

La méthode la plus simple pour trouver un document permettant de répondre à un besoin et de compter le nombre de fois où les mots de la requête apparaissent dans les documents, et de trier les résultats par ordre décroissant. Cette méthode n'est toutefois pas utilisée car les temps de traitement sont beaucoup trop grands. Un parcours complet de tous les documents à chaque requête est consommateur en ressources informatiques (calculs, mémoire). Néanmoins, le comptage des occurrences des mots de la requête reste le critère essentiel d'évaluation de la pertinence d'un document (voir paragraphe 2.3).

c) Indexation

Afin de résoudre le problème des temps de parcours des documents, et plus généralement, pour pouvoir faire correspondre une requête et un document quelconque (non textuel), l'index s'avère un moyen idéal. Un index est un tableau associant mots et documents. Pour chaque mot, il est possible d'obtenir la liste des documents qui le contiennent. L'index peut être enrichi de nombreuses informations comme le nombre d'occurrences d'un mot donné

[1] Répondre précisément à une question est un champs ouvert de l'intelligence artificielle : Les systèmes de question-réponse (Bellot, 2008)

dans un document, la distribution d'un mot dans un document, … La recherche se trouve désormais grandement simplifiée et accélérée, puisque le jeu de documents n'est plus parcouru. La difficulté du travail se reporte alors sur la construction de l'index et sur l'évaluation des résultats de recherche.

Le processus d'indexation est un parcours de tout nouveau document, une seule fois. Tout mot du document n'étant pas contenu dans l'index devient une nouvelle entrée. Toute entrée de l'index qui est contenue dans le document pointe désormais sur ledit document. Ce principe d'indexation est valable aussi bien dans un système d'information local (Système de fichiers sur un ordinateur, serveur documentaire) que sur le web. Dans ce dernier cas, les moteurs de recherche « envoient » des robots (petits programmes, *bot* ou *crawler* en anglais) chargés de répertorier le contenu des pages web. Les algorithmes d'indexation associés effectuent des pondérations selon des critères qui forment des secrets bien gardés par chaque acteur.

d) Modèle vectoriel

La forme la plus utilisée en RI pour l'index est le modèle vectoriel (Salton, 1971) et son amélioration LSI (Favre et. Al, 2006). Chaque document est représenté par un vecteur dont les indices sont les mots de l'index. La valeur de chaque dimension du vecteur est le nombre d'occurrence du mot correspondant dans le document.

	Mot 1	Mot 2	Mot 3	…
Document 1	0	2	1	…
Document 2	1	0	0	…
Document 3	0	0	3	…
…	…	…	…	…

Table 1 : La base documentaire est représentée sous la forme d'une matrice dans laquelle chaque document est un vecteur des nombres d'occurrences de chaque entrée de l'index.

La similarité entre documents, et la correspondance entre un document et une requête deviennent alors des mesures de distance entre vecteurs. La mesure la plus couramment utilisée est le cosinus. Il s'agit d'une mesure de différence entre les angles (les directions) des vecteurs pondérée par la taille de ces vecteurs. Deux documents proches sont représentés par deux vecteurs de

directions similaires (même sens). Cela s'applique de la même façon pour mesure la similarité entre un document et une requête (fig. 1).

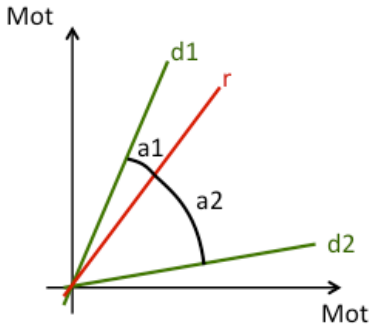

Figure 1 : Deux documents d1 et d2 sont plus ou moins similaires à une requête r dans le repère formé par deux mots (projection à deux dimensions).

Le processus de recherche revient mesurer la distance entre la requête et chaque document du système. La réponse (l'ensemble des documents retournés) du SRI est triée selon une mesure de pertinence, détaillé ci-après.

2.2. Métriques

Un des objectifs majeurs de la recherche d'information est de mesurer la qualité des réponses d'un système.

a) Qualité d'une réponse

La qualité d'une réponse sert à classer les résultats d'une requête. Cette qualité doit tenir compte de l'importance du mot dans un document et de l'importance relative de ce même mot dans l'ensemble des documents. Un mot apparaissant dans tous les documents n'est pas discriminant. La plus connue des mesures de pertinence est le TF.IDF.

TF (Term Frequency) : Nombre d'occurrences d'un mot donné dans un document, avec une pondération sur la taille du document. TF indique si le document traite du mot donné.

IDF (Inverse Document Frequency) : Inverse du nombre de documents contenant un mot donné. IDF précise si c'est un mot discriminant (rare dans l'ensemble des documents).

b) Qualité d'un SRI

En amont, les concepteurs d'un SRI doivent être capables de mesurer la qualité globale des réponses. En phase de test, ils simulent un jeu de requêtes sur un ensemble de documents dont ils ont au préalable évalué la pertinence pour chaque requête. Cela permet d'obtenir les taux de bonnes et mauvaises réponses et ainsi affiner les résultats.

Le taux de rappel d'un SRI est une mesure du silence. Il s'agit du nombre de documents pertinents retournés lors d'une recherche par rapport au nombre total de documents pertinents pour cette recherche. Un faible taux de rappel signifie que le SRI a oublié des résultats importants.

Le taux de précision est une mesure du bruit. Il s'agit du rapport entre le nombre de documents pertinents retournés lors d'une recherche et le nombre total de documents retournés. Un faible taux de précision signifie que de nombreux résultats ne sont pas pertinents.

2.3. Améliorations

Afin d'améliorer la qualité des résultats, il existe de nombreux moyens. Le premier d'entre eux est de ne pas considérer l'égalité stricte de deux mots. La fonction binaire (égale, non égal) utilisée initialement compare les lettres deux à deux. Un mot au singulier et son pluriel sont considérés comme différents. L'usage actuel veut que la reconnaissance des mots se fasse par similarité, la formule la plus connue étant celle de Levenstein. Elle consiste à compter le nombre d'opérations (ajout, suppression, déplacement de lettre) pour passer d'un mot à l'autre. Cette mesure n'étant pas spécifiquement adaptée à la langue, elle a été adaptée pour répondre à certains besoins spécifiques. Ainsi, le coût de transformation d'un « é » en « e » est moins important que pour d'autres lettres et permet de compenser certaines petites fautes d'orthographe. De même, la présence d'un « s » en fin de mot a un impact réduit et permet de considérer comme équivalent le singulier et le pluriel. Néanmoins, ce principe a tendance à apporter du bruit, et les réglages sont délicats et dépendants de la langue. Dans nos applications, nous avions par exemple trouvé un amalgame entre les mots « solaire » et « polaire ». Certains moteurs effectuent une analyse morpho-syntaxique pour affiner les ressemblances.

Une deuxième amélioration plus utilisée consiste à proposer à l'utilisateur une complétion de la requête en fonction des requêtes les plus effectuées. Très efficace dans un cadre général (sur le web), le principe à tendance à être pauvre à l'initialisation du système et trop fourni au fil du temps. En particulier dans les cadres restreints (SI d'entreprise), les requêtes autour du même thème vont

être systématiquement complétées de la même façon. Ce principe n'est d'aucune utilité pour les requêtes précises faisant référence à peu de documents.

2.4. Limitations des moteurs à mots-clés

Les moteurs de recherche à base de mots-clés fonctionnent relativement bien et offrent une précision assez intéressante. Ils possèdent par contre un défaut de rappel dans la mesure ou la recherche sur les mots ne permet pas de retrouver des documents traitant d'un thème sans utiliser les mots classiques. C'est tout le problème de l'utilisation de synonymes ou de figures de styles. Le cas de l'ellipse, très utilisée dans les domaines techniques, ne permet pas à un moteur de recherche de retrouver un document contenant « relais de tension » si l'utilisateur requière des documents contenant « relais à seuil » (<Relais à seuil de tension>)

Pour augmenter le taux de rappel sans perdre en précision, il est nécessaire de baser l'indexation et/ou la recherche sur une structure de données référençant les relations entre mots. C'est d'autant plus nécessaire si le moteur de recherche se veut multilingue.

2.5. Solutions intermédiaires

La solution du réseau lexical est souvent utilisée pour augmenter la qualité des moteurs, en particulier par la prise en compte de la synonymie. Wordnet en particulier est au cœur de projets d'indexation dite sémantique (Chevallet, 2008).

En marge, la folksonomie (Hotho, 2006) est un principe à la mode avec le caractère social que prennent les systèmes d'information et le web en général. Dans ce cas, ce sont les utilisateurs qui indexent les documents à la volée. Le principal avantage, c'est que cette indexation est le reflet d'une compréhension du document, et l'index ne contiendra pas nécessairement les mêmes mots. Le principal inconvénient est que l'indexation se fait par des points de vues particuliers dépendant d'usages particuliers. On trouve généralement des niveaux d'expertise variés qui décrivent les documents soient en termes généraux (pour le néophyte), soient en aspects précis (pour l'expert qui s'attache souvent à ce qui est considéré comme détail par le néophyte). A l'extrême, les communautés de pratiques différentes peuvent mener à des descriptions radicalement déconnectées. Les points de vues n'ayant parfois rien de commun, les résultats peuvent être surprenant.

Il est nécessaire de prendre en compte la compréhension d'un document dans la recherche d'information. La compréhension est le résultat d'un processus cognitif, spécifiquement humain. Elle mobilise des connaissances qui ne sont pas uniquement linguistiques. Nous pensons donc qu'il faut adosser le processus de recherche à une conceptualisation du domaine détachée des considérations linguistiques, tout en prenant en compte la langue qui reste le point d'accès aux documents. Notre solution passe par une formalisation ontoterminologique.

3. Ontoterminologie

3.1. Concepts et termes : systèmes sémiotiques distincts

« *L'ontoterminologie est une terminologie dont le système notionnel est une ontologie formelle* » (Roche, 2007). Elle est composée de deux systèmes sémiotiques distincts. D'une part, l'ontologie structure les concepts du domaine et fournit une compréhension indépendante de la langue. D'autre part, les termes fournissent les moyens d'exprimer les idées. La relation entre termes et concepts est décrite ci-après.

3.2. Méthodologie

La méthodologie de construction de l'ontoterminologie est discutée dans (Roche, 2007). Nous la rappelons ici pour insister le principe important de la séparation des systèmes de signes.

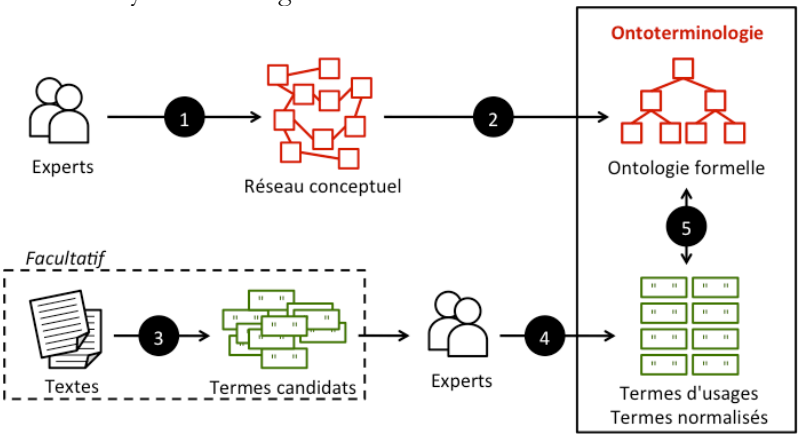

Figure 2 : Ontoterminologie : Une méthodologie mettant l'expert au cœur de du processus. Les concepts, en haut, sont séparés des termes, en bas. La difficulté du travail de modélisation réside dans la séparation des deux systèmes lors de leur construction.

La démarche se décompose en cinq parties révisables. La première vise à lister les concepts du domaine. Les experts sont au cœur de la démarche, interviewés et guidés par l'ingénieur cogniticien. Ils sont invités à définir précisément les objets sur lesquels ils travaillent. Rapidement, les informations se structurent sous la forme d'un réseau semi-formel. L'étape 2 consiste à formaliser ce premier résultat. L'ingénieur cogniticien, guidé par l'expert, rectifie le réseau selon une configuration formelle. Elle est ensuite révisée et validée. L'étape 3, facultative, utilise un extracteur de candidats termes pour lister une première série d'expressions faisant référence aux concepts. Les experts sont invités (étape 4) à fournir les termes utilisés dans le domaine par les différentes communautés de pratiques. Enfin, chaque terme est associé au concept qu'il désigne (étape 5).

3.3. Représentation, exemple

Afin de présenter la structure de l'ontoterminologie, nous nous appuyons sur des exemples tirés d'une ontologie qui traite de l'escalade. L'exemple présenté en figure 3 se focalise sur les différents ancrages. *Un ancrage sert au grimpeur à assurer sa sécurité. Il s'agit d'un dispositif d'assurage inséré, souvent fixé, dans le rocher auquel le grimpeur s'attache au fur et à mesure de sa progression.* L'exemple ne définit pas l'ancrage, mais les différentes sortes d'ancrage en fonction de la notion générale.

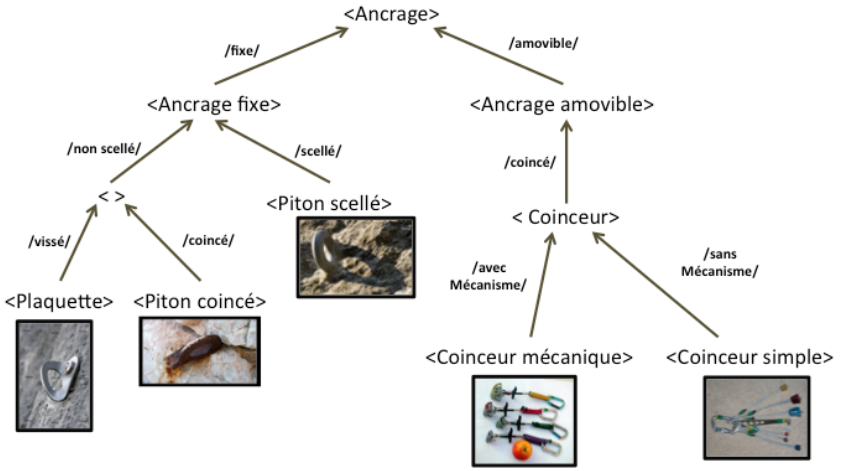

Figure 3 : Exemple partiel d'ontologie (simplifiée) : Les ancrage en escalade. Le terme « piton » en escalade désigne un piton coincé. Le <Piton scellé> n'est jamais appelé « piton » ni « piton scellé », mais plutôt « goujon ». Le terme « goujon » a quant à lui un sens plus général en mécanique.

Les concepts sont notés entre chevrons et nommés par une expression en majuscule. L'<Ancrage> est le concept parent et constitue la base de la définition formelle des concepts plus spécifiques. Les flèches représentent la relation de subsomption. Ainsi, <Ancrage fixe> est une sorte de <Ancrage>. La relation entre les deux concepts est étiquetée par une différence spécifique (Notation : encadrée par /) (Roche, 2001). La définition formelle de <Ancrage fixe> est <Ancrage> /fixe/ (*Un "ancrage fixe" est un "ancrage" qui a la caractéristique d'être "fixe"*). Cette définition semble apporter peu d'information car les expressions désignant les concepts s'incluent. La structure a ici plus d'importance que les mots utilisés, car c'est sur cette structure que les raisonnements s'appuieront.

L'arbre ontologique étends ses branches sur cette base jusqu'au niveau de précision requis. Les concepts les plus généraux (de premier niveau dans l'arbre ontologique) sont nommés « catégories ».

<Plaquette> $=_{def}$ <Ancrage>/fixe//non scellé//vissé/

<Coinceur mécanique> $=_{def}$ <Ancrage>/amovible//coincé//avec mécanisme/

La dimension terminologique regroupe l'ensemble des termes et les lie aux concepts. « goujon » est le mot généralement utilisé en escalade pour désigner un <Piton scellé> (<Ancrage>/fixe//scellé/), le nom du concept n'étant quasiment jamais employé. En mécanique, le terme « goujon » désigne une catégorie de matériel plus générale, incluant le goujon de l'escalade, d'où l'importance de définir formellement les concepts. Le terme « coinceur » n'est généralement pas utilisé pour désigner un <Coinceur>, mais plutôt <Coinceur simple>, le <Coinceur mécanique> étant plutôt appelé « friend »[2].

La figure 4 présente trois termes très utilisés pouvant apparaître dans le même texte, voir dans un même paragraphe. L'expression « huit » en escalade désigne autant un <Matériel d'assurage>, un <Nœud> qu'un <Niveau> de difficulté.

[2] Friend est une marque commerciale qui n'existe plus. Le terme « friend » est resté et désigne tous les coinceurs à mécanisme, quelque soit leur marque.

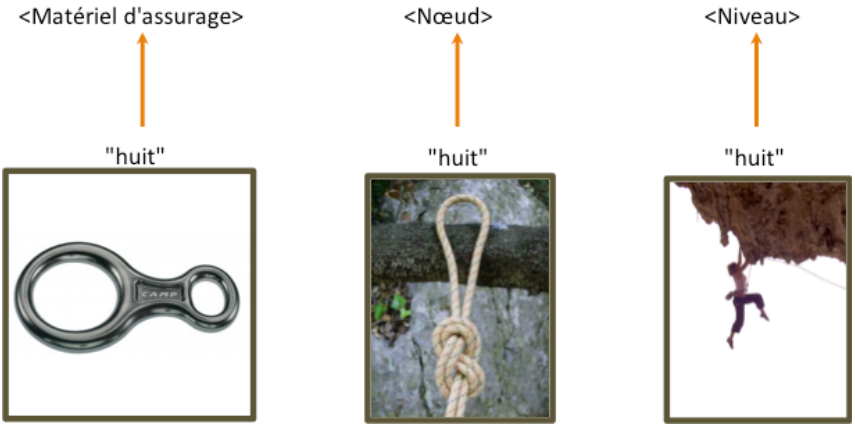

Figure 4 : Polysémie.

Le modèle ontoterminologique propose ainsi une architecture formelle à base de concepts associée à un répertoire de termes. Il est donc possible, d'une part d'effectuer des inférences, et d'autre part de faire référence aux contenus de documents. L'ontoterminologie constitue donc un outil idéal pour la recherche d'information sémantique.

La présentation du modèle n'est ici que partielle et ne conserve que certains fondamentaux nécessaires à la recherche d'information. L'ontoterminologie présente une dimension conceptuelle plus riche, avec des relations conceptuelles variées, ainsi qu'une dimension terminologique détaillée, avec définitions, catégorie grammaticales, réseau lexical…

4. RI Sémantique

La recherche d'information sémantique s'appuie en général sur une ressource extérieure de type thesaurus ou ontologie pour augmenter la qualité de la réponse. On dit que le SRI s'appuie sur une compréhension du domaine. Plus celle-ci est fine et validée, meilleurs sont les résultats. Nous présentons ici la RI sémantique telle que nous la concevons, basée sur l'ontoterminologie.

4.1. Marqueur sémantique

Comme dans la plupart des moteurs de recherche, notre proposition se base sur un index. Il est ici constitué de ce que nous nommons « marqueur sémantique ».

Le marqueur sémantique est la signature d'un document dans le contexte de l'ontoterminologie. Un même document peut avoir différents marqueurs sémantiques selon qu'il est considéré dans un domaine ou un autre. Produit de l'indexation, il contient les concepts identifiés dans le document, ou plus précisément, il est la liste des différences des concepts dont traite le document.

De manière plus formelle, un marqueur sémantique est un vecteur (tel que présenté au paragraphe 2.1) sur la base des différences spécifiques et des catégories de l'ontologie. Les valeurs de chaque dimension du vecteur décrivent la présence de la différence associée dans le document : valeur entre 0 et 1 (0 pour absent, 1 pour présent).

<Ancrage>	/fixe/	/Amovible/	/non scellé/	/scellé/	/vissé/	…
1	1	0	1	0	1	…

Table 2 : Partie de marqueur sémantique d'un document traitant de <Plaquette>. La définition complète du concept apparaît dans le marqueur.

La définition complète de chaque concept identifié apparaît dans le marqueur sémantique. Un document traitant d'un concept donné s'intéresse indirectement à ses concepts fils. Cette information est implicite dans le vecteur et pourra s'avérer utile pour la recherche.

4.2. Indexation

La construction du marqueur sémantique obéit aux mêmes règles générales que la construction d'un vecteur classique. Le document, généralement lemmatisé, est parcouru dans son intégralité. Chaque terme de l'ontoterminologie y apparaissant est identifié et les définitions des concepts associés sont ajoutées au marqueur sémantique. Si un terme est polysémique, le poids des différences dans le vecteur est divisé par le nombre de concepts les partageant. Ainsi, il est possible d'exprimer l'incertitude de présence d'un concept. Cette incertitude est relativement légère lorsqu'il s'agit de concepts parents, formulés dans le document à l'aide de métonymies. L'incertitude est plus lourde lorsque les termes sont « plus » polysémique (le nombre de différence partagées entre les concepts est faible). C'est le cas, à l'extrême, des trois versions du terme « huit » présenté plus haut. L'indexation est sujette à de nombreux paramètres et procédures que nous ne détaillons pas ici.

4.3. Recherche

La recherche en elle même ne présente pas de réelle différence par rapport à celle basée sur le modèle vectoriel initial (similarité vectorielle et tri). Les calculs

s'effectuent sur les marqueurs sémantiques. La requête est travaillée pour augmenter la quantité de résultats pertinents et diminuer le bruit. Dans un premier temps, les concepts correspondant aux termes de la requête sont extraits pour obtenir un marqueur sémantique identique à celui des documents, issu du même algorithme d'indexation.

La requête est implicitement simplifiée et augmentée sous la forme du marqueur sémantique, simplifiée car des redondances genre + espèce ne sont conservées que les espèces, augmentée car tous les documents contenant des termes désignant les concepts requis sont impliqués, y compris ceux ne comprenant pas les termes de la requête.

Ainsi, une requête de la forme « ancrage fixe coincé » sera signée par <Ancrage>/fixe//non scellé//coincé/ et sera interprétée naturellement comme traitant du concept spécifique <Piton coincé>. Les concepts encore plus spécifiques (possédant plus de différences) partageraient ces différences et seraient naturellement impliquée dans la recherche.

Ce travail sur la requête peut se résumer à deux actions principales :

- D'une part, tous les synonymes de tous les termes présents dans la requête sont ajoutés parce qu'ils sont autant de référence au concepts.

- D'autre part, tous les termes désignant des concepts plus spécifiques sont ajoutés, de manière à multiplier les résultats pertinents, même plus précis. Ceci est dû au partage des différences spécifiques. Ils ont en commun une partie de la définition.

Le classement des résultats par ordre décroissant de pertinence s'effectue selon une nouvelle mesure similaire au TF.IDF, mais actualisée pour la dimension conceptuelle. Le premier membre est la fréquence d'une concept "dans" un document : le nombre de fois où un des termes le désignant apparaît. Le second membre est l'inverse du nombre de documents où apparaissent les termes désignant le concept.

5. Meta-moteur

La section précédente présente le principe général de la recherche d'information appliquée à une base documentaire. La recherche d'information sémantique peut aussi être appliquée à Internet avec la différence majeure qu'il n'est pas possible d'indexer les documents.

Une solution consiste à développer ce qui est communément appelé un meta-moteur. Il s'agit d'utiliser les moteurs de recherche classiques et d'y injecter des aspects sémantiques. L'ontoterminologie peut être utilisée en amont sur la requête et en aval sur le traitement des résultats.

La requête de l'utilisateur est indexée comme précédemment (génération d'un marqueur sémantique). Les concepts requis sont identifiés et éventuellement désambiguïsés. Le système reformule la requête en utilisant tous les termes désignant les concepts identifiés. Ceci permet aux moteurs du web de retourner plus de documents, y compris des documents ne contenant pas les mots de la requête initiale. Une demande sur le mot « coinceur » effectuera aussi une recherche sur le mot « friend ». L'ajout de ces termes amène un bruit qui est supprimé par l'augmentation de la requête avec les différences dénomination du domaine. Ceci permet d'éliminer du bruit. Il est donc possible, par exemple, de retourner des documents traitant de « goujon » sans parler de mécanique. Enfin, les termes désignant les différents sous-concepts des concepts identifiés peuvent être recherchés afin de trouver des documents très spécifiques pouvant répondre au besoin de l'utilisateur. Ansi, une personne recherchant des informations sur les « pitons » peut être intéressée par des documents traitant de « plaquette ».

6. Conclusion, perspectives

La recherche par mot-clé présente des limites, même si l'habitude est bien ancrée et que les essais successifs permettent d'obtenir de bons résultats. Ceux-ci restent incertains et la qualité est difficilement vérifiable. La recherche sémantique s'impose dans des cadres techniques. L'ontoterminologie est un support efficace à la recherche d'information puisqu'elle permet de prendre simultanément en compte la compréhension du domaine et les mots pour en parler.

Notre approche a été validée et reste en cours de validation de deux manières. D'une part, nous avons mené diverses simulations sur des documents indexés manuellement pour vérifier que notre outil est efficace. Ces simulations devront s'accompagner d'une expérimentation réaliste pour un passage à l'échelle de la vérification expérimentale. D'autre part, l'outil est en usage dans diverses organisations et offre une satisfaction aux utilisateurs. Cette validation reste informelle, mais montre une certaine qualité de l'approche.

Malgré tout, il reste des évolutions possibles. Actuellement, le système ne prend en compte qu'une seule relation conceptuelle. Même si la relation

« sorte-de » reste à nos yeux capitale car elle représente la définition des concepts, d'autres types de relation devraient être pris en compte. La composition décrit des objets formés à partir d'autres objets. Chercher un document à propos d'un concept devrait dans un certaines mesure chercher ses composants. Il en va de même pour la relation de fonction.

Du point de vue de la structure du document, il existe des documents très pertinents à propos d'un concept ne contenant un terme le désignant que dans le titre. Avec les méthodes s'appuyant sur la fréquence des mots, ce document serait considéré comme non pertinent. Il faudrait ici accorder un poids lors de l'indexation à certaines parties du document. Cette tâche est délicate et requiert un certain nombre d'études, en particulier inter-langues.

Enfin, du point de vue de la langue, il restera toujours le biais des certaines figures de style, en particulier la métaphore qui va chercher les termes dans des domaines disjoints. Plus pragmatiquement, la lemmatisation des documents et la diversité des langues reste une barrière qui laisse ouvert le thème de la recherche d'information.

Bibliographie

Baeza-Yates R., Ribeiro-Neto B. (1999). *Modern Information Retrieval.* Ad- dison Wesley, New York, USA.

Bellot P., Boughanem M., "*Recherche d information et systèmes de questions-réponses*", 2008 in " La recherche d'informations précises : traitement automatique de la langue, apprentissage et connaissances pour les systèmes de question-réponse (Traité IC2, série Informatique et systèmes d'information)", sous la direction de B.Grau, Hermès-Lavoisier, chapitre 1, p. 5-

Chevallet J-P. (2009). "*Ressources endogènes et exogènes pour une indexation conceptuelle intermédia*", Habilitation à diriger des recherches

Favre B., Béchet F., Bellot P., Boudin F., El-bèze M., Gillard L., Lapalme G., Torres-Moreno J.-M. (2006) *The LIA-Thales summarization system at DUC-2006*, Document Understanding Conference (DUC-2006), New York (USA).

Hotho A., Jäschke R., Schmitz C., Stumme G. (2006) *Information Retrieval in Folksonomies:* Search and Ranking. ESWC 2006: 411-426

Roche C. (2001) : "T*he 'Specific-Difference' Principle : a Methodology for Building Consensual and Coherent Ontologies*", Actes de la conference IC-AI'2001, Las Vegas , USA

Roche C. (2008) : "*Le terme et le concept : fondements d'une ontoterminologie*", Actes de la première conférence TOTh 2007, Terminologie & Ontologie : Théories et applications, Christophe Roche éd., Annecy, Institut Porphyre, pp. 1-22, 2007

Salton G. (1971). T*he SMART Retrieval System—Experiments in Automatic Document Processing.* Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

A propos des auteurs

Damas Luc
Equipe Condillac, Université de Savoie
Campus Scientifique, 73370 Le Bourget-Du-Lac
Luc.damas@univ-savoie.fr
http://www.condillac.org
Tricot Christophe
Société Onomia
36 rue du clos d'Orléans, 94120 Fontenay-Sous-Bois
christophe.tricot@onomia.com

TOTh 2010. *Actes de la quatrième conférence TOTh - Annecy – 3 & 4 juin 2010*
Editeur : Institut Porphyre, *Savoir et Connaissance*
http://www.porphyre.org
Annecy, 2010
ISBN 978-2-9536168-1-1
EAN 9782953616811