Un traitement sémantique par ontologie pour l'indexation de documents dans un référentiel métier

Wilfried Njomgue Sado, Dominique Fontaine

Un traitement sémantique par ontologie pour l'indexation de documents dans un référentiel métier

Wilfried Njomgue Sado[1, 2], Dominique Fontaine[1]

[1] UMR CNRS 6599 Heudiasyc, Université Technologie de Compiègne,
BP 20529, F-60205 Compiègne
{wilfried.njomgue-sado, dominique.fontaine}@hds.utc.fr
[2] Suez Environnement CIRSEE Pôle Informatique Métier
38, rue du Président Wilson, F-78230 Le Pecq

Résumé : Cet article présente et évalue une approche sémantique qui a été greffée sur une approche originellement linguistico-statistique pour l'indexation de document. Elle combine en amont l'annotation sémantique du document à indexer via l'utilisation d'une ontologie de domaine, l'analyse linguistique du document et enfin l'analyse statistique par la décomposition en valeurs singulières des mots composant le document. Le système d'indexation qui a été développé sur cette base a pour tâche d'affecter tout nouveau document aux activités d'un référentiel métier préexistant. Nous en présentons les résultats, obtenus lors des diverses expérimentations menées sur un corpus de documents propre à la société Suez-Environnement.

Mots-clés : indexation, ontologie, représentation, gestion des connaissances

1 Introduction

Les spécialistes de la documentation assurent le stockage et la diffusion de l'information initialement fixée sur différents supports. Or, ces opérations exigent au préalable un traitement intellectuel des documents à savoir l'indexation. L'indexation en recherche d'information ne couvre pas seulement les aspects d'accès aux données mais aussi la représentation sous forme réduite d'un document par rapport à sa structure et à son contenu sémantique. On parle alors d'indexation par le contenu.

Cette problématique est celle du projet de gestion des connaissances, qui est initié par la Direction Technique et de Recherche de Suez-Environnement, et qui concerne tous ceux qui gèrent le patrimoine textuel du groupe. L'objectif général est d'accroître la valeur d'usage de l'Intranet du groupe qui est un outil clé de stockage, de partage et de diffusion d'information au sein de l'entreprise. Il est un réservoir de nombreuses connaissances (expérience et savoir-faire), généralement proposées sous forme de rapports et de notes techniques. Il favorise les échanges d'expériences entre les exploitants et des utilisateurs répartis sur l'ensemble des continents, en mettant à leur disposition des connaissances utiles à la réalisation de leurs activités.

Ce projet vise en particulier à faciliter l'accès aux documents qui supportent ces connaissances. Pour ce faire, un référentiel métier, une importante taxonomie qui décrit l'ensemble des activités ou métiers de l'entreprise, a été élaboré par des experts de l'entreprise. Il permet à l'utilisateur de sélectionner les documents qui l'intéressent. Il faut donc au préalable que ces documents aient été affectés judicieusement. Ce travail incombait jusqu'alors à des intervenants humains qui indexaient les documents de façon manuelle.

La masse de documents à introduire étant considérable, il a été entrepris d'automatiser ou plutôt de semi-automatiser le processus d'indexation. Nous avons alors conçu un système fondé sur une approche à la fois linguistique et statistique, qui assure l'affectation des nouveaux documents, après validation de l'utilisateur (Njomgue et Fontaine, 2004b). Nous avons ensuite fait l'hypothèse qu'il serait possible d'améliorer les résultats en adoptant au préalable une approche à caractère sémantique, objet principal de cet article.

Après une présentation du problème tel qu'il nous a été posé à l'origine, nous rappelons quelques caractéristiques de l'approche linguistico-statistique qui a d'abord soutenu notre système d'indexation. Ensuite, nous présentons les principes de l'approche à caractère sémantique qui, dans la version remaniée du système d'indexation, va finalement précéder le traitement linguistico-statistique. Celle-ci repose principalement sur la construction et l'utilisation d'une ontologie du domaine qu'il a fallu adapter à nos besoins spécifiques d'indexation. Nous faisons alors une comparaison des résultats obtenus avec ou sans traitement sémantique, sur une même collection de documents. Enfin, nous concluons sur quelques perspectives.

2 Problématique et processus

L'élément déterminant dont la présence influence et conditionne la totalité de notre démarche est le référentiel métier dont la connaissance nous est donnée a priori. Ce référentiel est actuellement une arborescence dont les feuilles sont les activités élémentaires du groupe dans le domaine de l'eau. L'auteur _ nom donné à celui qui introduit un nouveau document numérisé_ s'efforce de le classer en parcourant cette arborescence et en identifiant les activités qui selon lui le caractérisent au plus près. Cette tâche s'avère fort fastidieuse : en effet, l'auteur est d'abord censé connaître la plupart des activités de l'entreprise, hypothèse fort risquée, puis doit élaborer sa propre représentation du document, et enfin choisir certains métiers du référentiel parmi la multitude des possibilités. Il est donc essentiel de l'aider dans cette tâche en l'automatisant partiellement ou totalement, et si possible de réduire le temps nécessaire à son accomplissement.

Les particularités et les contraintes de la problématique sont alors les suivantes :

- l'indexation du document est faite relativement aux activités de l'entreprise, et non relativement aux mots du document. Le référentiel est donc à la fois une contrainte dans la mesure où il nous faut se conformer à un usage, à des pratiques, et aussi un support d'informations susceptibles d'orienter et d'aider l'indexation.

- nous ne sommes pas responsables de l'intégrité et de la pertinence du référentiel qui comporte des relations entre concepts dont la sémantique est pour le moins variable voire parfois indiscernable. En outre, il ne nous est pas permis de modifier cette structure parce qu'elle résulte d'un grand effort fait par la compagnie pour clarifier ses activités.
- le système doit être intégré dans un environnement Notes : cette intégration n'a pas été sans incidence sur certains choix techniques.
- nous sommes en mesure d'évaluer systématiquement les résultats produits par le système. En effet, nous les comparons à ceux fournis manuellement par l'auteur du document. Nous faisons l'hypothèse que les propositions faites par l'auteur sont pertinentes et donc qu'elles ne sont pas à remettre en cause. Cette contrainte est extrêmement forte car la diversité des auteurs fait qu'ils n'ont pas toujours la même compréhension du référentiel. Cet élément accroît la nécessité de concevoir un système semi-automatique, la décision revenant finalement à l'auteur.

On sait que l'indexation doit détecter les concepts caractérisant le mieux le document. Cette tâche est difficilement automatisable dans son intégralité, la plupart des systèmes n'indexant pas de façon totalement autonome les textes numérisés. L'indexation devient alors semi-automatique, l'intervention humaine étant garante de la qualité des résultats. C'est également le cas du système présenté, qui après examen du nouveau document propose une liste ordonnée d'affectations possibles à des feuilles du référentiel métier. L'auteur doit alors opter pour certaines d'entre elles.

Les systèmes d'indexation conçus ces dernières années empruntent des approches diverses : linguistique, statistique, plus rarement sémantique. En particulier, les méthodes mixtes apparaissent complémentaires et connaissent un succès notable, au vu des résultats positifs qu'elles fournissent : la meilleure d'entre elles semble être l'approche linguistico-statistique (LS) que nous avons aussi retenue pour notre processus d'indexation. Une première version de notre système a alors été élaborée par application d'une méthode LS.

Au fil des nombreuses expérimentations que nous avons menées, les résultats obtenus, répondant majoritairement à nos attentes (Njomgue et Fontaine, 2004a), ont certes permis de vérifier le bien-fondé de cette approche, mais ont aussi parfois révélé, ponctuellement, des insuffisances fâcheuses en termes de classification. Après analyse des cas considérés comme non conformes, nous avons estimé que la cause de ces défaillances résidait essentiellement dans l'absence de traitement ou de pré-traitement sémantique (Njomgue et Fontaine, 2004a), et même que, compte-tenu de l'existence du référentiel, la prise en compte d'informations à caractère sémantique s'imposait.

On remarquera à cet égard que la démarche sémantique sur la plupart des systèmes d'indexation se résume à un regroupement morphologique et/ou synonymique des termes clés extraits lors de la phase linguistique. Nous avons voulu aller plus loin dans l'analyse et le traitement sémantique des documents à indexer. Nous avons alors émis l'hypothèse que l'utilisation adéquate d'une ontologie du domaine, spécifiquement conçue pour l'indexation, pourrait être la clef de l'amélioration du système existant.

En définitive, le processus d'indexation de la version modifiée, considéré dans sa globalité, comporte plusieurs phases où s'enchaînent des traitements linguistiques, statistiques (Njomgue et Fontaine, 2004b) et donc plus récemment sémantiques. Le schéma descriptif de ce processus, commenté dans les ***sections 3 et 4***, est le suivant :

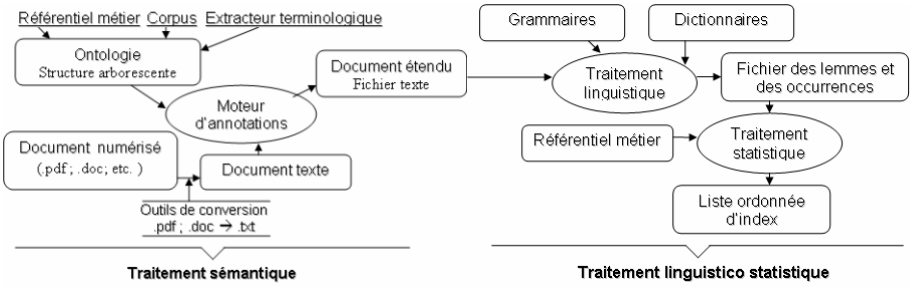

Fig. 1 – Schéma descriptif du processus d'indexation

On remarquera que, parmi les différentes options envisageables, notre choix fut de commencer par un traitement sémantique, dans le but de modifier le document afin d'en obtenir une représentation plus significative des concepts qui le caractérisent. La version résultante est alors soumise au traitement linguistico-statistique, qui pour ne pas biaiser la comparaison des méthodes a été strictement laissé dans la configuration de la version précédente du système.

Il va de soi que nous aurions pu faire un choix différent, par exemple celui qui aurait consisté à maintenir en premier le traitement linguistique et statistique pour seulement ensuite appliquer un traitement sémantique, cette fois-ci non plus dans le but de corriger le document, mais plutôt directement les résultats de l'indexation.

Le processus d'indexation linguistico-statistique ayant déjà été présenté par ailleurs (Njomgue et Fontaine, 2004b), nous en faisons une description simplifiée, puis nous décrivons les principaux aspects de la méthode de traitement sémantique, et en particulier ceux de l'ontologie sur laquelle elle s'appuie. On notera qu'à chaque phase de ce processus nous avons adopté une combinaison de différentes techniques, associées en parallèle ou en séquence. Les choix ont alors été fondé soit, *a priori*, sur des caractéristiques connues de ces techniques, soit, *a posteriori*, sur des résultats d'expériences.

3 Le traitement linguistico-statistique (LS)

Le traitement LS vise *in fine* à établir une correspondance entre le document d'origine et les activités du référentiel métier : il s'agit donc doublement de pointer les concepts représentatifs du document, (i.e.) d'accomplir une tâche d'indexation, et d'associer à cette forme indexée du document des feuilles du référentiel métier, (i.e.) d'effectuer une tâche de classification.

Le *document texte,* dénué d'images, vidéo, etc. est constitué uniquement des composantes textuelles du *document numérisé* initial Le *document étendu* est obtenu par enrichissement du *document texte*. Le traitement L.S. débute par une analyse linguistique du *document étendu,* qui consiste à extraire les termes composant le document, puis s'achève par une analyse statistique qui vise à mettre en valeur les termes importants.

L'analyse linguistique comprend séquentiellement des analyses morphologique et syntaxique, qui extraient et éventuellement regroupent les mots en fonction respectivement de leurs formes et de la syntaxe des phrases. Elle comprend également une étape dont la sémantique n'est pas totalement exclue, à savoir un regroupement synonymique autour termes clés du référentiel. Dans cette phase linguistique, sont notamment mises à contribution des grammaires locales, des dictionnaires spécialisés et des listes de synonymes relatifs aux termes du référentiel métier. Des techniques de lemmatisation (réduction automatique des mots à une forme de surface canonique), de stemming (réduction des formes de surfaces similaires à un seul concept), de stop-list (élimination de mots non pertinents pour l'indexation),en rapport avec le référentiel métier, sont mises en oeuvre.

L'analyse statistique qui est menée présuppose qu'il existe d'une part une relation entre la fréquence d'un terme dans un document et son importance pour ledit document, et d'autre part un lien entre l'importance d'un terme et le nombre de documents du corpus qui le contiennent. Elle discrimine les mots extraits à l'issue de la phase linguistique selon leurs occurrences et leurs co-occurrences, puis estime la proximité entre le document et les thèmes du référentiel.

A cet effet, de nombreuses techniques sont mises à contribution, parmi lesquelles, le latent semantic indexing ou LSI, la pondération des termes, en fonction de leur *contexte local* _par rapport au document _, de leur *contexte global* _par rapport à la base de données_ et de leur *contexte positionnel* _par rapport aux autres termes (leurs co-occurrences). Le processus statistique mis en oeuvre ici est une combinaison de diverses méthodes statistiques contribuant à l'indexation des documents, combinaison valable sous l'hypothèse qu'il est possible et pertinent d'associer les avantages des unes et des autres.

4 Traitement sémantique

4.1 Approche par enrichissement du document

Le système d'indexation doté d'une approche IS donne des résultats souvent satisfaisants (Njomgue et Fontaine, 2004b) comme nous le rappelons dans la section 3, mais fournit aussi des résultats parfois inadéquats, lorsqu'on les compare à ceux proposés par les auteurs des documents sur lesquels les expérimentations ont été conduites. Ces insuffisances s'expriment en termes de *silence*, lorsque les propositions d'affectation d'un document, émises par le système, ignorent les propositions privilégiées par l'auteur dudit document, ou de *bruit*, lorsque les

propositions d'affectations d'un document accordent une place prioritaire à des affectations que l'expert a lui-même jugées impropres.

L'analyse des cas problématiques nous a permis de discerner essentiellement trois situations qui en sont à l'origine : la présence de mots ou d'expressions ambigus, l'absence ou la sous représentation de certains mots ou de certaines expressions pourtant jugés importants, et *a contrario* la surreprésentation et donc la survalorisation de certains d'entre eux pourtant jugés peu pertinents. Ces situations évoquées ne peuvent alors qu'induire des distorsions sur l'ensemble des affectations proposées. Considérons deux exemples :

- supposons que le terme *"réseau"* apparaisse dans le document avec une forte occurrence, sans qu'il soit davantage qualifié explicitement. Il ne sera alors pas possible de discriminer entre des activités liées au *"réseau de distribution"*, au *"réseau informatique"* ou encore au *"réseau des eaux usées"*, thèmes appartenant au référentiel métier. Le terme *"réseau"* devient alors un index à la fois important et ambigu. Seule la prise en compte du contexte va permettre de lever cette ambiguïté, par exemple par la présence réitérée des expressions *"eau pluviale"*, *"eau parasite"* ou *"eau industrielle"* qui permettent sans guère de doute de rattacher le document aux activités liées au *"réseau des eaux usées"*.
- l'activité dénommée *"analyse des substances organoleptiques"*, explicitement décrite dans le référentiel, est importante dans le domaine de l'eau, et concerne beaucoup de documents à caractère technique, au point d'en devenir le sujet essentiel. Or cette expression est rarement présente sous cette forme ou sous une forme synonymique. En revanche, la présence de paramètres tels que le *"goût", "l'odeur", "le charbon actif", "la couleur", "la saveur",* etc. évoque irrésistiblement pour les spécialistes l'activité en question. Là encore, la prise en compte appropriée du contexte devrait permettre de revaloriser l'expression terme *"substance organoleptique"* et donc d'orienter l'indexation dans ce sens.

Notre problématique est alors la suivante : comment prendre en compte l'ensemble du document, à la fois les termes du référentiel explicitement décrits, mais aussi ceux qui ne sont que suggérés, connotés, déductibles ou obtenus par associations d'idées ? Comment introduire peu ou prou le contexte, lié au domaine, et en particulier les informations communes aux auteurs, absentes du document et pourtant décisives lors de la discrimination ? et enfin comment apporter un surcroît d'efficacité à l'approche LS dont on sait qu'elle accorde une place prééminente à la fréquence des occurrences ou des co-occurrences de mots ? L'approche que nous avons retenue repose sur deux principes :

- l'enrichissement du document par *l'adjonction* de mots ou d'expressions jugés, et ce dans le but de faire apparaître les concepts jusqu'alors manquants en ajustant leur fréquence d'apparition à la mesure de leur importance.
- et pour ce faire l'utilisation d'une *ontologie* du domaine dont on attend qu'elle nous apporte certains des liens conceptuels originellement absents

des documents, et en particulier ceux qui font référence aux concepts du référentiel métier.

En somme, le traitement sémantique, précédé d'un prétraitement linguistique minimal (conversion, mise au format texte, lemmatisation), consiste ici, par une *méthode d'enrichissement*, à refléter *quantitativement* l'importance *sémantique* et *qualitative* des concepts.

4.2 Edification de l'ontologie

En Ingénierie des Connaissances, les ontologies sont des conceptualisations d'un domaine (Gruber, 95 ; Fernandez Lopez, 1999; Asucion Gomez-Perez, 1999), pour nous ici celui des métiers de l'eau et en particulier de ceux portant sur l'analyse et le traitement de l'eau. Malgré leur diversité, un invariant est qu'elles contiennent des concepts et la spécification des relations entre ces concepts.

Les ontologies sont en général définies pour un objectif donné. Il en est ainsi pour l'ontologie utilisée par notre système puisqu'il s'agit finalement d'aider à l'indexation de documents. En conséquence, cette ontologie n'a en aucun cas un caractère générique, et sa réutilisation est largement dépendante du type d'application qui y fait appel.

Diverses approches ont été proposées pour édifier des ontologies (Aussenac-Gilles et al. 2000 ; Kassel G., 2002). N'ayant pas à disposition un unique expert pour l'ensemble des activités, nous avons retenu une approche de construction de l'ontologie *à partir de textes* en suivant le processus présenté en figure 2. Les phases ont été les suivantes :

- *la phase de constitution du corpus* est importante car elle est à la base de cette approche. Des experts nous ont fourni un panel de documents techniques représentatifs du domaine : notre *corpus numérisé*.

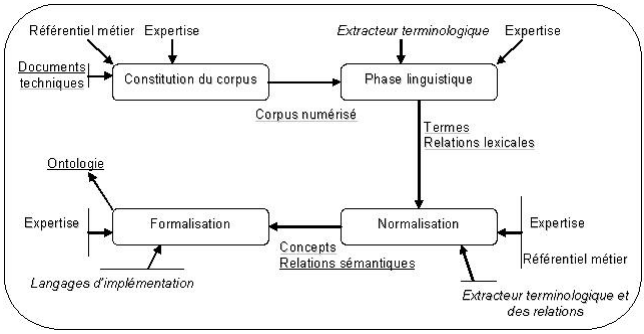

Fig. 2 – Edification d'une ontologie à partir des textes

- *l'étude linguistique* sur le corpus a été menée à l'aide d'extracteurs terminologiques, ici Intex (Silberztein, 2001) et Syntex (Bourigault, 2002). Nous avons notamment défini des grammaires de reconnaissance de certaines relations (synonymie, hyperonymie, hyponymie, etc.) (Aussenac-Gilles et Séguéla, 2000 ; Hamon et al. 1999) afin non seulement d'extraire les mots, les relations et de les désambiguïser, mais aussi de visualiser les relations dans

les contextes où elles apparaissent. Par ailleurs, nous nous sommes appuyés sur le référentiel métier (le corpus du métier), les critères de validation des concepts à la sortie des extracteurs reposant essentiellement sur l'importance que leur octroie le référentiel.

- *la normalisation* a consisté à interpréter sémantiquement et à structurer les termes à travers notre connaissance du métier, du corpus et des regroupements résultant de l'étude linguistique. Nous avons alors obtenu des concepts et des relations sémantiques, en attente d'une validation experte.
- *la formalisation* est la phase d'implémentation, d'élaboration et de validation de l'ontologie (Barry et al. 2001). Nous avons adopté une approche centrifuge (Gandon et Dieng, 2001 ; Faure et Poibeau, 2000) en identifiant les concepts clés du référentiel métier et en complétant l'ontologie par généralisation ou spécification des relations et concepts extraits du corpus. La dernière phase a consisté à implémenter l'ontologie avec Java, en raison en particulier de sa compatibilité avec l'environnement Notes de Suez.

4.3 Utilisation de l'ontologie pour l'indexation

4.3.1 Représentation de l'ontologie

La structure sémantique de l'ontologie du système d'indexation est composée de :

- *concepts :* ce sont des mots abstraits tel *"distribution "*, ou concrets tel *"eau pluviale"*, élémentaires ou composés. Nous aboutissons à près de 1200 concepts clés du domaine de l'eau (Lyonnaise des Eaux, 1994).
- *relations* qui représentent un type d'interaction entre les concepts du domaine (Guarino et al. 2001 ; Kassel, 2002 ; Studer et al. 1998 ; Zweigenbaum et Grabar, 2000). La version actuelle en comporte près de 1500.
  Les relations binaires utilisées entre concepts, largement majoritaires, sont de types subsomption (*"l'altrazine"* est_un *"pesticide"*), synonymie, mérony-mie, et lien morphologique (*"chloration*" a pour morphème *"chlore"*).
  D'autres sont éventuellement n-aires (n≥2) : ce sont les relations de causalité (*" traitement des eaux usées"* cause *"boue"*), évocation (*"l'indice de molhman"* évoque *"épaississement des boues"*). En outre, on remarquera que ces relations sont souvent *incertaines*, ce dont on tiendra compte, et que les relations d'évocation sont fortement liées à la tâche d'indexation.

Pour représenter ces différentes et nombreuses relations, nous avons opté pour la représentation unifiée par règles de production :

- *typées* de la forme P_1 *et ...* *et* P_k ® C, de prémisses P_i, de conclusion C, où la flèche est étiquetée par le type de la relation,
- et *pondérées* par un poids p accordé en fonction de la certitude accordée à la relation ontologique, égal à 1 si la relation est certaine sinon $0 < p < 1$.

Cette représentation, simple et largement éprouvée, nous a notamment permis de regrouper les règles en fonction de leurs types. De plus, les algorithmes développés pour les systèmes de production répondaient à nos besoins : pour enrichir automatiquement le document à indexer, il nous restait alors à les transformer en un *moteur d'annotations* de document.

4.3.2 Enrichissement du document

Cette phase de traitement sémantique par enrichissement du document, désormais partie intégrante du processus d'indexation sémantico linguistico statistique que nous noterons S-L-S, vise à mieux identifier le sujet du document analysé, par l'adjonction de concepts clés du domaine de l'eau.

Un *moteur d'annotations* à caractère déductif est dédié à cette tâche. Il ne considère *dans la version actuelle* que des fenêtres d'analyse circonscrites aux phrases. En effet, le but du traitement étant d'indexer le document en référence à une ou plusieurs idées fédératrices, nous avons estimé judicieux de focaliser l'attention sur les phrases : en effet, la circonscription à des phrases révèle des liens entre mots plus forts et plus chargés de sens que ne le révèlerait une recherche d'occurrences simultanées étendue à l'ensemble du document.

Les principes d'enrichissement sont alors inspirés de ceux qui régissent le comportement des moteurs d'inférences, pour les systèmes à base de règles de première génération. En balayant le document à indexer, le moteur d'annotations déniche les prémisses des différentes relations qui sont simultanément présentes au sein d'une même phrase. Lorsque tel est le cas, la conclusion des relations concernées est ajoutée au document. Il se peut bien sûr que le concept ajouté soit lui aussi prémisse d'autres relations ontologiques. Il importe donc de réitérer le cycle précédent, et ce jusqu'à épuisement des possibilités d'ajout. A titre d'exemple, *le "chlore" est un "désinfectant" ; le "désinfectant" a un lien morphologique avec la "désinfection" ; la "désinfection" est un "traitement de finition dans le traitement des eaux usées"* Ainsi, si "chlore" est un concept du document, "désinfectant" et "traitement de finition" seront ajoutés au document.

Parmi de nombreux choix possibles, le traitement de l'incertitude pour lequel nous avons opté, au départ à titre expérimental, est issu des méthodes à coefficients de vraisemblance en vigueur dans différents systèmes de première génération. Sans en détailler ici le mécanisme, il associe un principe de propagation de l'incertitude au sein d'un réseau de règles à un principe de renforcement de l'incertitude. En effet, si plusieurs relations permettent d'évoquer la présence d'un même concept chacune avec son propre facteur d'incertitude, il est légitime d'associer ces divers facteurs pour conclure à une pertinence accrue de ce concept. En définitive, le critère retenu pour l'inscription définitive d'un nouveau concept dans le document est qu'à l'issue de ce traitement celui-ci soit affecté d'un facteur de vraisemblance supérieur ou égal à 0.7, seuil qui expérience faite s'est révélé satisfaisant.

Par ailleurs, il est apparu que ce principe d'expansion du document par ajouts successifs de concepts clés devait être maîtrisé. En effet, si un concept apparaît *n* fois dans un document, alors la conclusion de la relation dont il est une prémisse sera

insérée *n* fois dans le cas d'une relation certaine, ces insertions étant répercutées sur ses descendants par propagation. Compte-tenu de l'objectif qui est de mieux prendre en compte les concepts à la fois importants et sous-représentés en augmentant leur fréquence d'apparition, cet effet multiplicateur était bien souhaité, mais en même temps il a fallu le contenir, sous peine de dérapage potentiel.

5 Expérimentations

Pour mener notre étude, assurer le développement du système d'indexation et enfin effectuer les expérimentations nécessaires au choix progressifs des méthodes et à la validation du système, nous avons disposé assez rapidement d'une part de documents issus de la société Suez-Environnement, écrits en français courant et souvent à caractère technique, et d'autre part des affectations au référentiel proposées par les auteurs de ces documents. Au final, nous avons effectué nos expérimentations sur un panel de près de 450 documents. L'hypothèse sous-jacente aux évaluations fut bien entendu que ces affectations de référence étaient pertinentes, même si l'analyse du contenu de certains documents nous a parfois laissé perplexes.

Généralement, les auteurs suggèrent manuellement au plus 10 activités pour un document. Le système d'indexation réduit et propose au maximum 15 activités parmi les 165 possibles, classées par ordre de pertinence, et parmi lesquelles en dernier lieu l'auteur doit choisir.

Nous avons constaté que la moyenne de mots utiles dans un document est d'environ 700 mots. Pour estimer la pertinence des indexations et ici des affectations, entre autres critères, il est usuel d'évaluer les performances par le rappel et la précision qui désignent respectivement le nombre de documents pertinents retournés par rapport au nombre total de documents pertinents et le nombre de documents pertinents retournés sur le nombre total de documents retournés.. De plus, compte tenu du caractère semi-automatique du système, il est primordial de limiter autant que faire se peut le silence, et il est plus pertinent d'évaluer le système par le rappel que par la précision. A l'issue de ces expériences, les résultats sont les suivants (Table 1) :

Table 1. Résultats des expérimentations

		Sans traitement sémantique		Avec traitement sémantique	
	Index (auteur)	Mots outils	Rappel index (système)	Mots outils	Rappel index (système)
Minimum	1	35	0%	73	0%
Maximum	10	1537	100%	1603	100%
Moyenne	4	700	77.09%	800	87%

A ce stade, et notamment au vu des réactions des premiers utilisateurs de l'entreprise, ces résultats sont apparus conformes à nos espérances : pour la méthode S-L-S, un taux de rappel fort satisfaisant, et sur les parties où l'ontologie a été validée,

aucun cas n'a donné lieu à des réponses erratiques. En outre, et tel était l'objet de cet article, il ont permis de confirmer lintérêt qu'il y avait à ajouter un traitement sémantique au système, et à le faire par référence à une ontologie. L'adjonction de ce traitement a permis en particulier de diminuer très sensiblement le silence, ce qui était notre première préoccupation, et incidemment et indirectement de réduire le bruit en reléguant plus loin dans la liste les propositions inappropriées.

6 Conclusion et perspectives

Dans une première étape, un traitement sémantique du document permet d'annoter, d'ajouter du sens aux documents par rapport au référentiel métier du groupe et au domaine de l'eau. La méthode présentée s'appuie sur une ontologie du domaine qui a été édifiée en vue de servir à l'indexation. Ensuite, le traitement linguistique représente le document à indexer par un ensemble de concepts jugés lexicalement significatifs. En dernière étape, un traitement statistique discrimine ces concepts en considérant leurs occurrences et leurs co-occurrences, puis estime la proximité entre le document et les activités cibles du référentiel.

Globalement, les résultats obtenus sont à considérer en rapport avec la difficulté de la tâche : un référentiel assez hétérogène et imposé, une grande diversité de documents, une ontologie non stabilisée sur certains aspects faute d'expert désigné, et un jeu de tests dont la pertinence ne semblait pas toujours indiscutable.

L'évaluation comparative des méthodes, linguistico-statistique (LS) seule ou précédée d'un traitement sémantique (S-L-S), confirme le bien fondé de l'apport de la sémantique dans le domaine de l'indexation des documents à forte composante textuelle, et plaide en faveur de l'utilisation d'ontologies du domaine pour l'indexation.

Cette évaluation fait en même temps apparaître quelques lacunes en terme de précision, surtout compte tenu des exigences de qualité qu'ont légitimement les futurs utilisateurs de ce système. Nous nous efforçons de pallier ces insuffisances, notamment en affinant les mécanismes d'enrichissement des documents, par exemple par un meilleur contrôle de l'effet multiplicateur ou une meilleure utilisation de la typologie des relations. Il s'agit là d'une problématique de recherche qui nous semble porteuse de perspectives intéressantes. Nous pensons également que la prise en compte de la structure des documents permettra d'effectuer l'enrichissement avec plus de sélectivité.

Par ailleurs, nous cherchons à stabiliser l'ontologie du domaine, et à gommer les points faibles. Les experts et auteurs sont aujourd'hui sollicités dans cette phase d'amélioration et au-delà d'exploitation de l'ontologie. Cet effort doit aboutir à terme à un système intégré dans un environnement Intranet.

Références

ASUCION GOMEZ-PEREZ. (1999). Développements récents en matière de conception, de maintenance et d'utilisation d'ontologies. Revue n°19 : Terminologie et Intelligence Artificielle

(Actes du colloque de Nantes, 10-11 Mai 1999) de RINT : Réseau International de Néologie et de Terminologie.

AUSSENAC-GILLES N. , BIEBOW B. ET SZULMAN S. (2000). Modélisation du domaine par une méthode fondée sur l'analyse de corpus. In Ingénierie des Connaissances 2000

AUSSENAC-GILLES N. ET SEGUELA P. (2000). Les relations sémantiques : du linguistique au formel.

BARRY C., CORMIER C., KASSEL G. ET NOBECOURT J. (2001). Evaluation de langages opérationnels de représentations d'ontologies. Actes de la conférence IC'2001. Pages 309-327. Grenoble. France

BOURIGAULT D. (2002). Upery : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus, Actes de la 9ème conférence annuelle sur le Traitement Automatique des Langues (TALN 2002), Nancy, 2002, pp. 75-84

GANDON F., DIENG R. (2001). Ontologie pour un système multi-agents dédié à la mémoire d'entreprise. Actes de la conférence IC'2001, Pages 1-20. Grenoble. France

FAURE, D. POIBEAU T. (2000). First experiments of using semantic knowledge learned by ASIUM for information extraction task using INTEX. In Staab, S. and Maedche, A. and Nedellec, C. and Wiemer-Hastings P., editors, Ontology Learning ECAI-2000 Workshop, pages 7-12.

GUARINO N., GANGEMI A., MASOLO C. ET OLTRAMARI A. (2001). Understanding top-level ontological distinctions. Workshop on Basic Ontological Issues in Knowledge Sharing. IJCAI'95.

GRUBER T. R. (1995). Towards Principles for the Design of Ontologies Used for Knowledge Sharing. International Journal Of Human Computer Studies. 43 : 907-928. Stanford University

HAMON T., GARCIA D., NAZARENKO A. (1999). Détection de liens de synonymie : complémentarité des ressources générales et spécialisées. Revue n°19 : Terminologie et Intelligence Artificielle (Actes du colloque de Nantes, 10-11 Mai 1999) de RINT : Réseau International de Néologie et de Terminologie.

KASSEL G. (2002). OntoSpec : une méthode de spécification semi-formelle d'ontologies. Actes des 13ièmes journées francophones d'Ingénierie des Connaissances, IC' 2002. Pages 75-87. France

LYONNAISE DES EAUX (1994). Mémento du gestion de l'alimentation en eau et de l'assainissement. Tome 1, 2 et 3.

NJOMGUE W., FONTAINE D. (2004A). A linguistic and statistical approach for extracting knowledge from documents. TAKMA 2004 (Fifth International Workshop on Theory and Applications of Knowledge MAnagement) in Conjunction with DEXA 2004 (Database and Expert Systems Applications); Spain

NJOMGUE W., FONTAINE D. (2004B). Identification des thèmes d'un document relativement à un référentiel métier. In Manifestation de JEunes Chercheurs Sciences et Techonologies de l'Information et de la Communication. France

SILBERZTEIN, M. (2001). Intex @ manual ASSTRIL - LADL, 201p, 2000-2001.

STUDER R., BENJAMINS V. R., FENSEL D. (1998). Knowledge Engineering : Principles and Methods. Data & Knowledge Engineering. 25 : 161-197.

FERNANDEZ LOPEZ, M. (1999). Overview of Methodologies for Building Ontologies, The Proceedings of the IJCAI-99 Workshop on Ontologies and Problem Solving Methods, Sweden August-1999.

ZWEIGENBAUM P., GRABAR N. (2000). Liens morphologiques et structuration de terminologie", in IC 2000 : Ingénierie des connaissances, pp. 325-334.