De la linguistique aux statistiques pour indexer des 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

Mots-clefs – Keywords

Linguistique, indexation, recherche d'information, statistique

Linguistics, statistics, indexing, information processing

Résumé – Abstract

Cet article présente une méthode d'indexation automatique de documents basée sur une approche linguistique et statistique. Cette dernière est une combinaison séquentielle de l'analyse linguistique du document à indexer par l'extraction des termes significatifs du document et de l'analyse statistique par la décomposition en valeurs singulières des mots composant le document. La pondération des termes tire avantage de leur contexte local, par rapport au document, global, par rapport à la base de données, et de leur position par rapport aux autres termes, les co-occurrences. Le système d'indexation présenté fait des propositions d'affectations du document à un référentiel métier dont les thèmes sont prédéfinis. Nous présentons les résultats de l'expérimentation de ce système menée sur un corpus des pôles métiers de la société Suez-Environnement.

This article presents an automatic method of documents indexing based on a hybrid, linguistic statistical approach. The proposed approach combines a linguistic analysis of the document by the extraction of the significant terms of the document in conformity with the referential; and a statistical analysis of the same document decomposed into separed words. Innovating weighting of terms is set to take judiciously advantage of both their position with respect to other terms (co-occurrence) and their local and global context. An application was developed in order to assign referential-based topics to documents. Finally, we will present experiments results and evaluation carried out on documents of Suez-Environnement Company.

1 Introduction

Les opérations de stockage et la diffusion des documents sur différents supports exigent au préalable une indexation qui consiste à réduire le contenu sémantique de chaque document. La Direction Technique et de Recherche de Suez-Environnement a initié un projet de gestion des connaissances dont l'objectif principal est de concevoir un outil qui permette à tout utilisateur d'introduire de nouveaux documents dans la base de données de l'Intranet du groupe. La particularité de ce projet réside dans l'existence d'un référentiel métier qui a été élaboré il y a quelques années et est constamment mis à jour. Il s'agit d'une taxonomie qui décrit l'ensemble des activités et métiers de l'entreprise. Avant de mettre un document dans la base de donnée, son auteur doit identifier au mieux le sujet du document, en fonction des activités qui en caractérisent la sémantique. Cette tâche d'indexation s'avère fort fastidieuse : en effet, l'auteur est d'abord censé connaître la plupart des activités de l'entreprise, hypothèse fort risquée, puis élaborer sa propre représentation du document, et enfin choisir certains métiers du référentiel parmi la multitude des possibilités. Il est donc essentiel de réduire le temps nécessaire à l'accomplissement de cette tâche, en l'automatisant partiellement ou totalement. La plupart des systèmes n'indexent pas de façon totalement autonome les textes numérisés, on parle alors d'indexation semi-automatique. Notre système propose une liste ordonnée d'affectations possibles du nouveau document à des métiers du référentiel afin que l'auteur puisse opter pour une ou plusieurs d'entre elles.

Le présent article a pour objectif principal de présenter un processus d'indexation qui permet d'analyser chaque document et de déterminer les affectations possibles en fonction des métiers du référentiel. Il présente d'abord la méthode d'indexation automatique, ensuite évalue la méthode sur une collection de documents, et conclut sur quelques perspectives.

2 Une méthode d'indexation automatique

L'indexation présentée ici a pour but d'extraire les concepts identifiant au mieux le document, puis de le rattacher aux métiers prédéfinis au sein du référentiel. L'indexation du document est faite relativement aux activités de l'entreprise, et non relativement aux mots du document. La contrainte supplémentaire est la suivante : nous ne sommes pas responsables de l'intégrité et de la pertinence de ce référentiel qui comporte des relations entre concepts dont la sémantique est pour le moins variable voire au pire indiscernable. En outre, il ne nous est pas permis de modifier cette structure. 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 tout en faisant 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, d'où la nécessité de concevoir un système semi-automatique.

Pour accomplir cette tâche, le processus d'indexation, considéré dans sa globalité, s'appuie à différents moments sur le référentiel, et comporte trois phases principales, où s'enchaînent successivement des traitements linguistiques, statistiques et sémantiques. Nous n'abordons dans cet article que les deux premières phases, la phase de traitement sémantique, basée sur l'exploitation d'une ontologie du domaine, étant encore en cours de finalisation.

3 Analyse et traitement linguistique

Il s'agit ici d'extraire automatiquement les termes composant un document (Séguéla, 2001 ; Bourigault, Jacquemin, 2000). Parmi la multitude d'outils ayant de très bonnes performances (Ana, Termino, Syntex, Intex, Acabit, etc.), Intex (Silberztein, 2001) a retenu notre attention car il permet d'intégrer des dictionnaires spécialisés, des grammaires de reconnaissance des syntagmes, répondant en particulier à nos besoins. Le traitement linguistique comprend alors séquentiellement des analyses morphologique, puis syntaxique, et sémantique, celui-ci se réduisant à un regroupement morphologique et/ou synonymique des termes clés). Dans le but de compléter l'analyse linguistique, nous avons inséré un dictionnaire spécialisé, et des grammaires locales afin de détecter, modifier ou réduire certaines abréviations afin de ne pas perdre l'information associée aux abréviations. Pour réduire la taille des mots non lemmatisés, nous avons construit des grammaires de reconnaissances de certains lemmes. Des grammaires de corrections de certains mots erronés ont également été mises sur pied afin de ne pas biaiser l'information du document.

Au terme de cette étape, nous obtenons deux fichiers texte : un fichier «taggé » F_{tag} , fichier où les lemmes non ambigus sont écrits entre accolades et un fichier F_{lemme} , fichier des termes lemmatisés et des occurrences associées. Afin d'affiner les lemmes ainsi obtenus, nous appliquons premièrement un « stemming » sur le fichier F_{lemme} pour obtenir le fichier que nous noterons $F_{Stemming}$. Le stemming est la réduction des formes de surfaces similaires à un seul concept, par exemple « inintéressant », « intéressant », « intérêt », « intéressé » seront réduits en « intérêt ». Ce « stemming » s'est basé sur la liste des mots du référentiel métier ayant préalablement subie un « stemming » manuel. L'objectif de cette méthode est de donner plus d'importance aux termes métiers du domaine. Ensuite, nous obtenons le fichier issu de la phase linguistique par application de la technique dite de « stop-list » sur le fichier $F_{Stemming}$ (Table 1). Celle-ci consiste souvent à dresser une liste de termes non soumis à l'indexation. Ici, plutôt que d'utiliser une liste prédéterminée, nous avons fait le choix d'écarter les termes dont le nombre de caractères est inférieur à une valeur fixée empiriquement (3 est la valeur optimale obtenue par expérience), considérant que ce sont des termes de poids sémantiquement faibles (« le », « la », etc.), et donc peu intéressants à être indexés.

Table 1. Résultats des applications des techniques de « stemming » et de « stop-list »

Fichier des lemmes F_{Lemme}	Fichier « stemming » $F_{Stemming}$	Fichier linguistique
{S} 3 {activation} 1 {activer} 1 {ainsi} 1 {aire} 1 {alimentation} 2 {analyser} 1 {analyse} 2	{S} 3 **{activer} 2** {ainsi} 1 {aire} 1 {alimentation} 2 **{analyse} 3**	 {activer} 2 {aire} 1 {alimentation} 2 {analyse} 3

4 Un traitement statistique

Nous présentons le processus statistique à travers successivement la méthode de pondération des termes existantes dans le document, la recherche des proximités entre les termes fondée sur la notion de co-occurrence, et enfin l'application de la méthode du « latent semantic indexing ».

4.1 Méthode de pondération

Pour mettre en valeur un terme par rapport à un autre, le système le pondère. Avant de procéder au choix de notre modèle de pondération, il nous a paru utile de faire une rapide synthèse des différentes méthodes existantes. De toutes les pondérations existantes, (Singhal et al., 1996) affirment que la pondération $P_{1\alpha}$ est la plus intéressante parmi celles ne prenant pas en compte la composante globale pour la recherche d'information. Pour celles qui prennent en compte cette composante, $P_{2\alpha}$ est intéressante d'après (Faraj et al., 1996). La composante globale est le facteur qui permet d'accorder un poids plus important aux termes discriminants qui apparaissent moins fréquemment dans la collection des documents.

Nous noterons par $P_{f\alpha}$ le poids du terme α du profil lexical P_{lex} ; $C_{i,j}$ le poids de co-occurrence du couple des termes (u_i, u_j) ; P_i le poids du terme u_i dans le document ; n_c le nombre de fois où les termes u_i, et u_j apparaissent ensemble ; n_i le nombre de fois où le terme u_i apparaît seul, N le nombre de documents dans la base de donnée ; tf_i l'occurrence du mot i dans un document ; df_j le nombre de documents dans lequel le terme j apparaît.

$$P_{1\alpha} = [1+\log(tf_\alpha)] \times \left[\frac{1}{\sqrt{\sum_{\alpha=1}^{n}[1+\log(tf_\alpha)]^2}}\right] \quad \text{et} \quad P_{2\alpha} = [1+\log(tf_\alpha)] \times \left[\log\left(\frac{N}{df_\alpha}\right)\right]$$

Nous définissons alors, à partir de ces deux pondérations, une méthode de pondération par étude de cas (Table 2). Cette pondération est normalisée dans [0,1]. On dira que la valeur de la pondération est respectivement faible, moyenne et forte si cette valeur est incluse dans [0 ;0.25],] 0.25 ; 0.75[, [0.75 ; 1]. Tous les termes du fichier linguistique sont soumis à cette méthode.

Table 2. Pondération définitive des termes. $\text{Max}(P_{1\alpha}, P_{2\alpha})$ et $\text{Min}(P_{1\alpha}, P_{2\alpha})$ désignent respectivement le maximum et le minimum entre les pondérations $P_{1\alpha}$ et $P_{2\alpha}$.

Pondération définitive	**Pondération $P_{2\alpha}$: elle met plus en avant la composante globale que ne le fait $P_{1\alpha}$**			
Pondération $P_{1\alpha}$: elle met plus en avant la normalisation que ne le fait $P_{2\alpha}$		Forte	Moyenne	Faible
	Forte	$\text{Max}(P_{1\alpha}, P_{2\alpha})$	$P_{1\alpha}$	$P_{1\alpha}$
	Moyenne	$P_{1\alpha}$	$P_{1\alpha}$	$P_{1\alpha}$
	Faible	$P_{2\alpha}$	$P_{1\alpha}$	$\text{Min}(P_{1\alpha}, P_{2\alpha})$

4.2 Approche matricielle : adaptation du « Latent Semantic indexing »

Plusieurs applications dans le domaine de la recherche d'information (RI), de la classification des documents, du filtrage d'information (Deerwester et al., 1990 ; Dumais et al., 1996) ont été développées selon l'approche matricielle du « Latent Semantic Indexing » (LSI) qui forunit de meilleurs résultats par rapport aux méthodes standards. Ici, on suppose qu'il y a une structure « latente », à caractère « sémantique », dans l'usage des mots d'un document qu'on révèlera par la décomposition en valeur singulière du LSI. Le plus souvent, la matrice (« unités lexicales » x « unités textuelles ») est le point de départ de cette méthode. Dans le cas présenté, nous utilisons respectivement les thèmes de ce référentiel et les termes du profil lexical pour définir les unités

textuelles et les unités lexicales et obtenir ainsi une matrice Termes-Thèmes notée $X_{Termes,Thèmes}=(X_{i,j})$. Ainsi, la sémantique d'un document est considérée comme une combinaison linéaire (Dumais et al., 1996) du contenu des thèmes du domaine ainsi que du sens des termes associés. Notre but étant d'affecter un document par son contenu au référentiel, il est pertinent de représenter un document en liaison avec les thèmes du domaine, un thème étant un chemin de l'arborescence qu'est le référentiel. Alors :

$$X_{i,j} = \begin{cases} P_i + \sum_{k=i+1}^{n} C_{i,k} & \text{si } Terme_i \subseteq Thème_j \\ 0 & \text{sinon} \end{cases} \quad \text{avec} \quad C_{i,j} = (P_i + P_j) \times Proxi(u_i, u_j)$$

$$Proxi(u_i, u_j) = \frac{n_c(P_i + P_j)}{n_i P_i + n_j P_j} \; avec \; n_c \leq \min(n_i, n_j)$$

$X_{i,j}$ est la contribution du terme i du document au thème j, relativement au document à indexer. Les colonnes de cette matrice représentent la distribution du sens de chaque thème pour le document. Tout document est une combinaison linéaire de la contribution sémantique des thèmes représentant le domaine. Après décomposition en valeurs singulières (Husbands et al., 1996), la matrice réduite contient seulement les premiers composants linéaires indépendants k de $X_{Termes\ -\ Thèmes}$ avec $\sigma_1 \geq ... \geq \sigma_k > 0$. Nous chiffrons numériquement chaque thème afin d'en extraire les plus représentatifs, en calculant la norme des colonnes de la matrice projetée des thèmes obtenue.

5 Expérimentations

Pour réaliser les expériences, nous disposons d'emblée de documents issus du groupe écrits en langage naturel libre : le français ; et des affectations au référentiel proposées par les auteurs de ces documents. Dans notre expérience, nous avons à notre disposition un ensemble de près de 450 documents. Il y a 165 thèmes possibles pour l'indexation d'un document. L'auteur fait un choix parmi les 15 thèmes proposés par le système semi-automatique. Il s'avère, dans cette étude, que la moyenne de mots utiles dans un document est d'environ 700 mots. Le système d'indexation étant semi-automatique, le silence, (i.e.) le fait que le système n'extraie pas un thème suggéré par l'auteur, nous préoccupe davantage que le bruit. Il nous est alors apparu judicieux d'évaluer le système (Table 3) en terme de rappel plutôt que de précision.

Table 3. Résultats de l'expérience : rappel

	Mots importants	Thèmes proposés par l'auteur	Rappel (système)
Minimum	35	1	0%
Maximum	1537	10	100%
Moyenne	700	4	77.09%

Le *rappel* est le nombre de documents pertinents retournés par rapport au nombre total de documents pertinents.

A ce stade, ces résultats sont plutôt satisfaisants (77,09% de rappel) eu égard à la difficulté de la tâche : un référentiel assez hétérogène et imposé, une grande diversité de documents, et un jeu de tests dont la pertinence n'est pas toujours avérée mais auquel il faut se conformer. Bien sûr dans l'absolu, quelques problèmes de silence demeurent, notamment dans les cas où aucun terme du

thème n'apparaît dans le document, où le document est trop technique (présence de vidéo au détriment du texte), et enfin où l'information au sein du document est implicite.

6 Conclusion et perspectives

Le bien fondé de l'approche mixte à savoir linguistique puis statistique est confirmé à travers cette évaluation. D'abord, l'utilisation des grammaires, des techniques de lemmatisation, de stemming, de stop-list permet de réduire le document à indexer à un ensemble de mots jugés intéressants. Ensuite, un traitement statistique discrimine ces mots en raison de leurs occurrences et de leurs co-occurrences, puis estime la proximité entre le document et les thèmes du référentiel par le biais du LSI.

Cette évaluation révèle quelques insuffisances en terme de précision et de silence. Ces insuffisances ont été signalées dès le début du projet et nous nous proposons de les résorber certes par des améliorations au traitement effectué, mais surtout par un traitement sémantique. A cet effet, nous mettons en place une ontologie du domaine de l'eau, spécialisée de façon à répondre à nos besoins en matière d'indexation. Les experts et auteurs de l'entreprise sont actuellement sollicités dans cette phase de construction et au-delà d'exploitation de l'ontologie.

Références

BOURIGAULT D., JACQUEMIN C. (2000): « Construction des ressources terminologiques, In Ingénierie des langues », pp 215-230, 2000, ed. J.M. Pierrel. Hermes Sciences

DEERWESTER S., DUMAIS S., FURNAS G., LANDAUER T., HARSHMAN R. (1990): « Indexing by Latent Semantic Analysis », *Journal of Society for Information Science*, Vol.41, n.6, pp. 391-407, 1990

DUMAIS S., LETSCHE T., LITTMAN M., LANDAUER T. (1996): « Automatic Cross-Language Retrieval using Latent Semantic Indexing », *SigIR Multilingual IR Workshop*, Aug. 22, 1996

FARAJ N., GODIN R., MISSAOUI R., DAVID S., PLANTE P. (1996): « Analyse d'une méthode d'indexation automatique basée sur une analyse syntaxique de texte », *Canadian Journal of Information and Library Science / Revue l'information et de bibliothéconomie*, 1996

HUSBANDS P., SIMON H., DING H. (1996): « On the use of Singular Value Decomposition for Text Retrieval », *Proceeding's of SIAM Comp. Information Retrieval Workshop, 2000.*

SEGUELA P. (2001): Construction de modèles de connaissances par analyse linguistique de relations lexicales dans les documents techniques, 2001, Université Toulouse III, France

SINGHAL A., SALTON G., BUCKLEY C. (1996): « Length normalization in degraded text collections », *Proceedings of Fifth Annual Symposium on Document Analysis and Information Retrieval*, April 15-17, 1996, pp. 149-162.

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