Création d'un espace conceptuel par analyse de données contextuelles

Nicolas Kumps, Pascal Francq, Alain Delchambre

CAD/CAM – Université Libre de Bruxelles – Av. F. D. Roosevelt, 50 – CP 165/14 1050 Bruxelles – Belgique
{nkumps, pfrancq, adelch}@ulb.ac.be

Abstract

The computerized treatment of numerical information requires a system of representation. The objective of the method presented in this paper is the entirely automated generation of a valid and compact system of representation of the information contained in documents. One hopes to group words referring itself on the same subject, the same idea in order to use these "concepts" to model information rather than to use the representation of these concepts.
The algorithm proposed generates a vector space of the concepts by gathering dimensions of a vector space of the terms. This regrouping is carried out by comparison of the parameters related to the context of appearance of the terms in the treated documents. The reduction of dimension of the space used will allow a faster data processing and requiring a less important memory capacity. This type of improvement being typically that required within the framework of real system, tallies of the application for which the method was developed.

Résumé

Le traitement informatisé de l'information numérique requiert un système de représentation. L'objectif de la méthode présentée dans cet article est la génération entièrement automatisée d'un système de représentation, valide et compact, d'informations contenues dans des documents. On espère grouper des mots se rapportant à un même sujet, une même idée afin d'utiliser ces « concepts » pour modéliser l'information plutôt que d'utiliser la représentation de ces concepts.
L'algorithme proposé génère un espace vectoriel de concepts en regroupant les dimensions d'un espace vectoriel de termes. Ce regroupement s'effectue par comparaison des paramètres liés au contexte d'apparition des termes dans les documents traités. La réduction de dimension de l'espace utilisé permettra un traitement de l'information plus rapide et demandant un espace mémoire moins important. Ce type d'amélioration étant typiquement celle recherchée dans le cadre de système réel, cadre de l'application pour laquelle la méthode a été développée.

Mots-clés : espace conceptuel, association de mots, similarité, cooccurrence.

1. Introduction

Ce travail s'inscrit dans le cadre d'un projet de recherche documentaire : le projet GALILEI[1]. Chaque utilisateur du système est représenté sous la forme d'un profil. Il pourra se définir autant de profils qu'il possède de centres d'intérêt. Il pourra alors juger les documents, au moyen d'un simple clique de souris, comme pertinent ou hors sujet par rapport à un de ces profils. Ceux-ci sont ensuite groupés en communautés virtuelles sur base des documents jugés. Un échange de documents pertinents est alors possible au sein d'une même communauté. Le fruit de la recherche documentaire de tous est ainsi partagé entre les utilisateurs du système (Francq, 2003).

[1] Subventionné par la Région wallonne sous le contrat 01/1/4675 – http ://www.galilei.ulb.ac.be

Comme pour tout projet de recherche documentaire un mode de représentation des documents est nécessaire. C'est donc dans l'optique d'une description de profil pour l'application de groupement que la méthode présentée dans cet article a été développée.

L'objectif de celle-ci est la substitution de l'espace vectoriel des termes par un espace vectoriel des concepts plus compact.

2. État de l'art

L'augmentation continue de la masse d'informations numérique a rendu son traitement informatisé inévitable. Pour le rendre possible, différents modèles de représentation des textes (vectorielle, probabiliste, connexionniste, ...) adaptés au mode de traitement informatique ont été développés. À titre d'exemple, on peut citer le modèle vectoriel standard (Salton et McGill, 1983), les modèles LSI (Deerwester *et al.*, 1990), SOM (Self Organizing Map),...

Pour permettre l'utilisation d'un modèle, l'indexation des documents est nécessaire. Celle-ci introduit toute une série de problèmes. Pour tenter d'y répondre, de nombreuses méthodes ont été développées. La segmentation d'un texte en unité linguistiques peut s'effectuer à l'aide de règles (tokenisation) ou sur base d'un lexique (lexematisation). Toutefois certains séparateurs ambigus — comme les mots composés, les synonymes ou encore les acronymes — rendent la tâche ardue et nécessite un grand nombre de traitements supplémentaires. Ainsi les ambiguïtés lexicales et sémantiques peuvent être en grande partie résolues par des méthodes d'étiquetage morpho-syntaxique (Brill, 1992 ; Church, 1988 ; Cutting *et al.*, 1992 ; Roche et Schabes, 1995) et de désambiguïsation sémantique (Krovetz et Croft, 1992), la variabilité des mots par des techniques de désuffixation (Paternostre *et al.*, 2002 ; Porter, 1980), l'utilisation d'un dictionnaire ou l'utilisation de règles de dérivation fondées sur une analyse linguistique (Tzoukermann et Jacquemin, 1997).

Il existe aussi différentes techniques permettant de prendre en compte l'importance relative des unités linguistiques. L'utilisation d'anti-dictionnaires permet l'élimination des termes nécessaires à la construction de la phrase, mais généralement sans contenu informationnel. L'extraction de l'information la plus pertinente peut être réalisée à l'aide de filtres fréquentiels ou encore par l'utilisation des informations apportées par une analyse syntaxique.

Parmi l'ensemble de toutes ces méthodes des choix doivent être effectués en fonction des objectifs recherchés.

3. Représentation de l'information dans GALILEI

Les impératifs de ce projet sont ceux d'un système réel, avec toutes les limitations qui y sont associées : temps de calcul, place mémoire, autant de paramètres à prendre en compte et à minimiser. C'est dans ce cadre que la méthode a été développée à partir de la représentation déjà disponible.

Les documents analysés se décomposent en unités linguistiques plus ou moins porteuses de sens. Afin d'alléger la quantité de termes différents à stocker, des traitements préalables ont été appliqués.

Certains mots courants, nécessaires à la construction de la phrase, n'apportent aucune information sur la nature du sujet traité par le document. En absence de traitement morpho-syntaxique, ces mots peuvent être supprimés par l'utilisation d'un anti-dictionnaire (*stoplist*). Un deuxième traitement est appliqué afin de supprimer la trop grande variabilité des mots. En effet les variabilités flexionnelles (pluriel, conjugaison) et les variabilités dérivationnelles (passage d'une

catégorie morpho-syntaxique à une autre) introduisent un grand nombre de termes différents raccordés à une même racine et donc à un même sens. Les techniques de désuffixation permettent de supprimer pour une bonne part ces variations morphologiques (algorithme de Porter pour l'anglais (Porter, 1980), Carry pour le français (Paternostre *et al.*, 2002)). Le dictionnaire formé suite à cette phase d'analyse sera donc composé de radicaux.

4. Méthode

La méthode présentée dans cet article a pour objectif de regrouper plusieurs termes significatifs en entitées représentatives que l'on qualifie de « concepts ». Cette association pourra être envisagée par la seule connaissance du contenu d'un ensemble de documents et devra permettre une réduction de l'espace de représentation tout en gardant le caractère discriminant de celle-ci.

Cette opération sera réalisée sur base de relations de similitude entre les différents termes au niveau de leur représentation dans les documents et de leur contexte d'apparition, mais aussi de leur spécificité et de leur capacité de description. La réalisation d'un espace conceptuel selon la méthode présentée peut se diviser en plusieurs étapes.

Dans un premier temps, les termes du dictionnaire susceptibles d'être groupés sont sélectionnés afin d'alléger le traitement (diminution du temps de calcul et de la capacité de stockage nécessaire). Sur base de ceux-ci l'algorithme peut être appliqué et des concepts sont ainsi formés. Enfin la transition vers le nouvel espace de représentation doit être effectuée pour les différents types d'information tels que les documents. Les différentes étapes sont donc :

1. La sélection des termes utilisés par l'algorithme.
2. La création des concepts.
3. La représentation de l'information sur base du nouvel espace créé.

4.1. *Sélection des termes*

En complément de l'anti-dictionnaire et des algorithmes de désuffixation, un filtre fréquentiel est utilisé pour ne garder que les termes les plus discriminants avant de leur appliquer l'algorithme proprement dit. Ce filtre éliminera ainsi les termes présents dans un nombre de documents trop faible ou trop élevé (ces nombres resteront proportionnels au cas étudié : nombre de documents et de catégories. Dans le cas des tests réalisés, ces limites sont de l'ordre de [10,300]).

4.2. *Création des concepts*

La création des concepts est réalisée par regroupement des unités linguistiques. Ce regroupement est effectué sur base de la comparaison des différentes caractéristiques de chaque couple de termes. Parmi ceux-ci, les plus proches sont rassemblés pour former les groupes.

Dans un premier temps, les termes sélectionnés sont classés selon l'ordre décroissant de leur facteur IDF (Inverse Document Frequency).

$$IDF_i = log\frac{|D|}{|D_i|} \tag{1}$$

$|D|$ = Nombre total de documents.
$|D_i|$ = Nombre total de documents contenant le terme i.

Cet indice permet de déterminer la capacité de chaque terme à représenter un petit groupe de documents. En effectuant un classement selon l'ordre décroissant, les premiers traités seront

ceux catégorisant le mieux l'information. Cela permet d'éviter de grouper en priorité les termes plus généraux présents dans de nombreux documents.

Ils seront ensuite comparés à tous les éléments de la liste et, sur base de leurs similitudes et des groupes déjà créés, insérés dans un groupe.

Les différentes caractéristiques à la base de la comparaison sont les suivantes :

1. Confiance ;
2. Similarité de voisinage ;
3. Cooccurrence.

4.2.1. *Confiance*

La confiance (Maedche et Staab, 2000) d'un terme au sein d'un couple exprime le nombre de documents dans lesquels les deux éléments du couple sont présents par rapport au nombre de documents dans lesquels ce terme apparaît.

$$confiance(i_i^k \Rightarrow i_j^k) = \frac{|\{d_k | i_i^k \cup i_j^k \subseteq d_k\}|}{|\{d_i | i_i^k \subseteq d_i\}|} \tag{2}$$

avec i_i^k et i_j^k, les deux termes composant le couple k.

À chaque terme est associé un certain nombre de documents, ainsi la confiance par rapport à un terme permet de décrire la part de ses documents décrite par un couple l'incluant. Chaque couple sera représenté par sa confiance minimum ainsi que par sa confiance moyenne.

$$confiance_{min} = min\{confiance(i_i^k \Rightarrow i_j^k), confiance(i_j^k \Rightarrow i_i^k)\} \tag{3}$$

$$confiance_{moy} = \frac{1}{2} \cdot (confiance(i_i^k \Rightarrow i_j^k) + confiance(i_j^k \Rightarrow i_i^k)) \tag{4}$$

4.2.2. *Similarité de voisinage*

Cette mesure se base sur la comparaison des contextes d'apparition de deux termes. Pour cela, à chaque terme i est associé un vecteur $\vec{V_i}$ des termes voisins du premier à travers tous les documents décrits par celui-ci. Ce vecteur est obtenu par déplacement d'une fenêtre de taille k à travers l'ensemble des documents analysés. La mesure de la similarité peut alors être obtenue par le produit scalaire normé des vecteurs de voisins des deux termes.

$$sim_v(i_i, i_j) = \frac{\vec{V_i} \cdot \vec{V_j}}{\|\vec{V_i}\| \cdot \|\vec{V_j}\|} \tag{5}$$

Cette mesure correspond au calcul de la matrice d'association scalaire à la différence que celle-ci ne se compose plus de tous les termes du dictionnaire (sélection) et que le voisinage comparé n'a plus la taille du document, mais celui d'une fenêtre de dimension k. Signalons encore que lors de l'analyse du document au moyen de la fenêtre de taille k tous les termes sont encore présents, lors de la construction du vecteur de voisinage ce n'est plus le cas. Ainsi chaque occurrence du terme n'apportera pas (k-1) voisins, mais bien

$$\{v_i\} \setminus \{\{v_i \in stoplist\} \cup \{v_i | |D_i| \notin [min, max]\}\} \tag{6}$$

où D_i est l'ensemble de tous les documents dans lesquels apparaît le terme v_i.

4.2.3. *Cooccurrence*

La cooccurrence de deux termes correspond au nombre d'apparitions simultanées de ceux-ci dans l'ensemble des documents, ou plus simplement le nombre de documents contenant les deux termes. Elle s'obtient en effectuant le produit scalaire des vecteurs documents de chacun des deux termes.

4.2.4. *Comparaison*

Il est évident que dans une certaine mesure ces différents facteurs sont liés (ex : deux termes ayant une cooccurrence nulle auront de la même manière des indices de confiance nuls). Toutefois chacune de ces mesures a sa propre spécificité. Il est en effet inutile de grouper deux termes n'apparaissant jamais simultanément dans un même document (cooccurrence) tout en n'étant pas exclu que bien qu'apparaissant peu dans les mêmes documents, ils soient relatifs à un même concept (similarité de voisinage). De la même façon, la confiance permet de se rendre compte du manque de précision de la mesure de cooccurrence car certains couples de cooccurrence élevée peuvent obtenir une confiance de l'ordre de 100 % pour un terme et ridiculement basse pour l'autre, indiquant par là la réunion d'un terme spécifique avec un terme très général. L'utilisation simultanée de ces différentes informations doit pouvoir permettre de remédier aux différentes difficultés rencontrées.

Le couple retenu sera celui répondant à la condition suivante :

$$\max_j \{confiance_{min}(i_i, i_j) + confiance_{moy}(i_i, i_j) + sim_v(i_i, i_j)$$

$$+(confiance_{min}(i_i, i_j) \cdot confiance_{moy}(i_i, i_j) \cdot sim_v(i_i, i_j))\} \quad (7)$$

avec des valeurs minimum pour la $confiance_{min}$ et la $cooccurrence$.

Ainsi, après formation et ordonnancement de la liste des termes à grouper, celle-ci est traitée itérativement terme après terme dans l'ordre de la liste. À chaque itération, le terme est comparé à tous les autres et est soit inséré dans un groupe déjà existant, soit un nouveau groupe est créé, suivant le résultat de la comparaison. Si aucun couple ne satisfait les conditions, le terme est supprimé de la liste.

4.3. *Représentation de l'information*

Le nouvel espace conceptuel étant formé, la translation d'un espace vectoriel à un autre peut être réalisée. Les documents ne seront plus représentés que par les concepts qu'ils contiennent. Un concept sera considéré comme apparaissant dans un document lorsqu'un minimum de deux des termes qu'il représente sont contenus dans le document. La fréquence associée à ce concept sera la somme de l'ensemble des termes associés au concept apparaissant dans le document.

5. Tests

Différents testes ont été effectué sur différentes bases de données pré-catégorisées : News1500 (anglais, 1480 documents répartis en 13 catégories), News20000 (anglais, 19778 documents répartis en 24 catégories), LeSoir (français, 13234 documents répartis en 9 catégories). Des bases de données pré-catégorisées sont utilisées pour permettre une division en profils et la simulation des jugements.

5.1. *Mesures*

L'objectif est de représenter l'information de manière plus compacte afin de diminuer l'espace de stockage et le temps de calcul. Les premières mesures à vérifier sont donc la variation du nombre de dimensions de l'espace de représentation et la qualité de la représentation.

Tant dans l'espace vectoriel des mots que dans l'espace des « concepts », les documents sont représentés par un vecteur $\vec{d_j} = (d_{1,j}, d_{2,j}, ..., d_{t,j})$ dont les composantes sont calculées de la manière suivante (Salton et McGill, 1983) :

$$d_{i,j} = \frac{o_{i,j}}{\max_l o_{l,j}} \cdot log\frac{|D|}{|D_i|} \tag{8}$$

$o_{i,j}$ = occurrence du terme i dans le document j.

La similarité entre deux documents sera calculée en effectuant le produits scalaire normé de deux vecteurs (similarité cosinus)

$$sim(\vec{d_i}, \vec{d_j}) = \frac{\vec{d_i} \cdot \vec{d_j}}{\|\vec{d_i}\|\|\vec{d_j}\|} \tag{9}$$

Similarité interne moyenne : La similarité interne est la similarité du vecteur d'un document par rapport au vecteur moyen de la catégorie à laquelle il appartient.

$$\frac{1}{|D|} \cdot \sum_{d_i \in D} sim(\vec{d_i}, \vec{d}^{k_i}) \tag{10}$$

$$\vec{d}^{k_i} = \frac{1}{|D_{k_i}|} \cdot \sum_{d_i \in D_{k_i}} \vec{d_i} \tag{11}$$

$\vec{d}^{k_i}$= vecteur moyen de la catégorie k_i à laquelle appartient le document i.

Similarité externe moyenne : La similarité externe est la similarité du vecteur d'un document par rapport au vecteur moyen de la catégorie à laquelle il n'appartient pas et dont il est le plus proche.

$$\frac{1}{|D|} \cdot \sum_{d_i \in D} \max_{j \neq i} \{sim(\vec{d_i}, \vec{d}^{k_j})\} \tag{12}$$

Pourcentage de documents bien catégorisés : Pourcentage de documents dont la similarité interne est plus élevée que la similarité externe. Ce sont donc les documents pour lesquels le vecteur est le plus proche du vecteur moyen de la catégorie à laquelle ils appartiennent.

Ces différentes mesures permettrons de vérifier que les différents sujets sont bien identifiés et que les profils pourront ainsi être groupés.

5.2. *Simulation*

L'objectif étant d'intégrer la méthode dans GALILEI, des simulations ont été effectuées sur les différentes bases de données. Pour simuler un système réel où les profils sont groupés en communautés virtuelles, la procédure utilisée est décrite ci-après.

On part d'une base de données catégorisée. Deux profils attachés chacun à une des catégories sont créés. Un ensemble de jugements sur les documents est simulé : un document appartenant à la même catégorie que le profil avec lequel il est jugé, est considéré comme pertinent, dans l'autre cas, comme hors sujet. Ils sont alors groupés en communautés virtuelles et ce résultat est comparé à la solution idéale.

Ensuite, se succèdent au fil du temps des phases de jugement des documents partagés au sein des communautés virtuelles et des phases de création d'un nouveau profil associé à un ensemble de jugements de documents.

Le test porte sur 500 unités de temps. Chaque unité de temps correspondant à une phase de jugement ou une phase de création de nouveau profil. À la fin de chaque unité de temps un nouveau groupement est calculé et la solution réévaluée. L'évaluation de la solution se fait à travers l'*Adjusted RAND Index* (Hubert et Arabie, 1985) qui représente la qualité de la réponse par rapport au groupement idéal. Cette valeur est comprise entre -1 et 1. Lorsque celle-ci est nulle, le groupement calculé est aléatoire.

6. Résultats

6.1. *Exemples de mots groupés (LeSoir) :*

douma - kremlin
medical - infirmier - hopital
gerhard - schröder
asterix - goscinny - uderzo
afghanistan - taliban - kaboul
échappement - rejet - consommation - pollution
...

On peut voir à partir de ces quelques exemples de groupements réalisés pour la base de données LeSoir, la cohérence des groupes créés.

6.2. *Mesure de similarité*

Data Base	langue	Représentation	similarité interne	similarité externe	pourcentage	dimension
News1500	en	mots	0.38302946	0.24871015	86.0135	38914
		concepts	0.37456777	0.21177723	88.0405	2304
News20000	en	mots	0.28735075	0.21038529	71.9486	109965
		concepts	0.19984695	0.08734833	74.6890	5590
LeSoir	fr	mots	0.60287311	0.60384188	60.7520	100471
		concepts	0.18022144	0.08921644	85.3786	5821

Tableau 1. Résultats

Les résultats permettent de constater une nette diminution de la taille de l'espace de représentation (rapport supérieur à 15) tout en apportant une différentiation entre catégories légèrement meilleure (+2 ou 3% pour l'anglais et +25% pour le francais).

6.3. *Simulation*

Les résultats des simulations sur les différentes bases de données montrent bien qu'à partir d'une certaine masse d'informations le système fonctionne de manière équivalente dans les deux

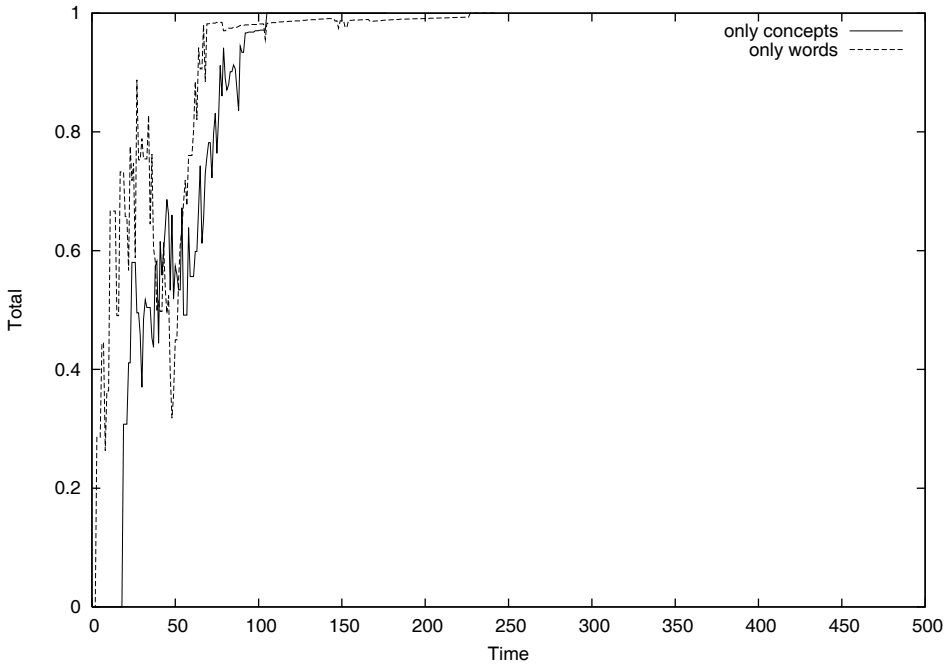

Figure 1. Simulation d'un système réel sur News1500

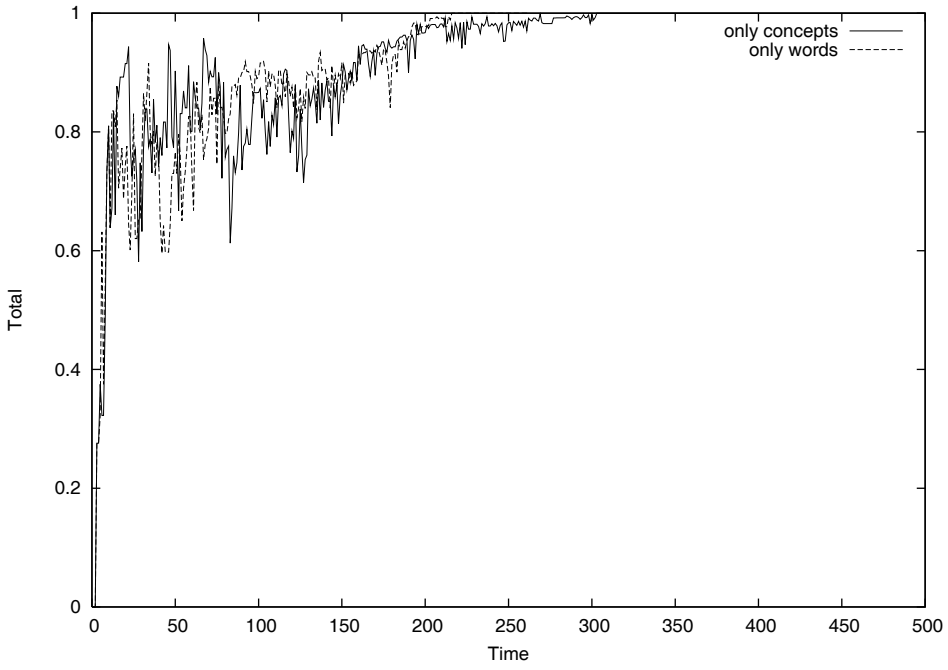

Figure 2. Simulation d'un système réel sur News20000

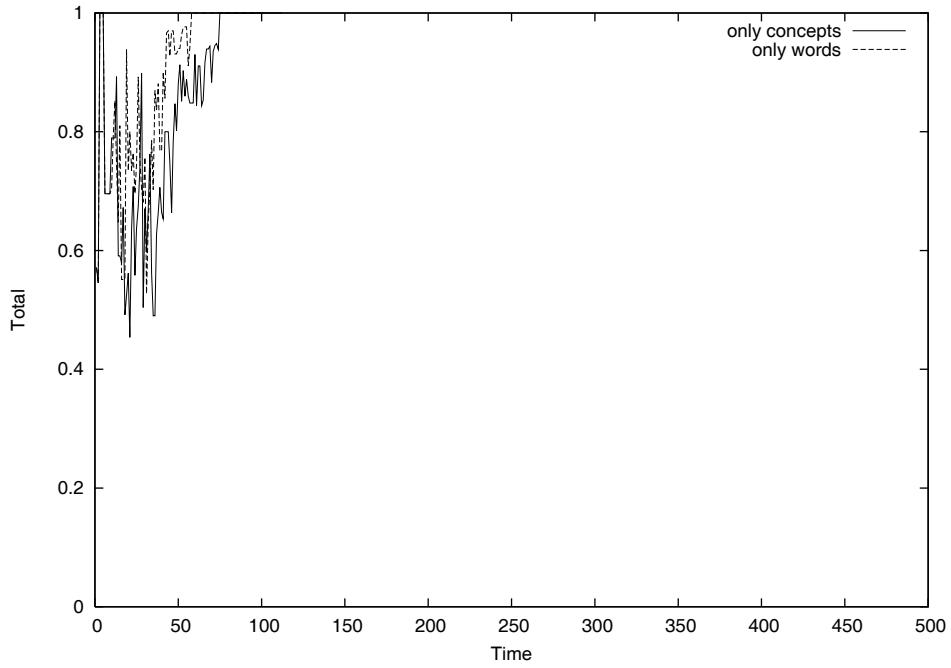

Figure 3. Simulation d'un système réel sur LeSoir

modes de représentation à la différence près que dans l'espace conceptuel le temps nécessaire au calcul est nettement moindre. (Figure 1, Figure 2, Figure 3)

7. Conclusion

L'objectif de la méthode proposée était la génération automatique d'un système de représentation valide et compact, dans l'optique d'un traitement de l'information plus rapide et moins gourmand en ressources. Les résultats obtenus montrent une diminution conséquente des dimensions sans perte importante d'information. La réduction des dimensions de l'espace de représentation sur base de l'analyse des données contextuelles a donc bien été réalisée.

L'utilisation d'une telle méthode permet donc un gain de temps relativement important au niveau des différents traitements appliqués à l'information ainsi qu'une diminution de la capacité de stockage nécessaire, tout en n'atténuant pas la qualité de la représentation. De plus, un de ces principaux avantages est son entière automatisation.

Références

Brill E. (1992). A Simple Rule-Based Part-of-Speech Tagger. In *Proceedings of ANLP-92, 3rd Conference on Applied Natural Language Processing* : 152-155.

Church K.W. (1988). A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text. In *Proceedings of the 2nd ACL Conference on Applied Natural Language Processing* : 136-143.

Cutting D., Kupiec J., Pedersen J. et Sibun P. (1992). A Practical Part-of-Speech Tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing* : 133-140.

Deerwester S., Dumais S., Landauer T., Furnas G. et Harshman R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society of Information Science*, vol. (41/6) : 391-407.

Francq P. (2003). *Stuctured and Collaborative Search : An integrated approach to share documents among users*. PhD Thesis. Université Libre de Bruxelles.

Hubert L. et Arabie P. (1985). Comparing partitions. *Journal of Classification* : 193-218.

Krovetz R. et Croft W.B. (1992). Lexical ambiguity and information retieval. *Information Systems*, vol. (10/2) : 115-141.

Maedche A. et Staab S. (2000). Discovering conceptual relations from text. In *Proceedings of the 13th European Conference on Artificial Intelligence* : 321-325.

Paternostre M., Francq P., Saerens M., Lamoral J. et Wartel D. (2002). Carry, un algorithme de désuffixation pour le français. Version électronique disponible sur [http ://www.galilei.ulb.ac.be].

Porter M.F. (1980). An Algorithm for Suffix Stripping. *Program*, vol. (14/3) : 130-137.

Roche E. et Schabes Y. (1995). Deterministic Part-of-Speech Tagging with Finite-State Transducers. *Computational Linguistics*, vol. (21/2) : 227-253.

Salton G. et McGill M. (1983). *Modern Information Retrieval*. McGraw-Hill Book Co.

Tzoukermann E. et Jacquemin C. (1997). Analyse Automatique de la Morphologie Dérivationnelle et Filtrage de Mots Possibles. In *Actes du Colloque Mots possibles et mots existants* : 251-260.