Recherche d'Information efficace utilisant la sémantique : le focus

Anthony Ventresque

Laboratoire d'Informatique de Nantes Atlantique (LINA)
2 rue de la Houssinière
BP 92208
44322 NANTES cedex 3
Anthony.Ventresque@univ-nantes.fr

RÉSUMÉ. *L'indexation sémantique de documents à partir d'ontologies est un domaine qui prend de l'essor, malgré les difficultés d'une indexation automatique ou même semi-automatique, sans parler d'indexation manuelle. Il est possible désormais d'avoir des caractérisations sémantiques de documents textuels ou non textuels basées sur des ontologies. Partant de ce fait, nous avons mis en place un objet, le focus, qui représente un document ou une requête en pondérant les concepts d'une ontologie de manière à indiquer l'importance de chaque concept dans le document. Néanmoins, pour être utilisable, cette solution doit définir précisément ce qu'est une normalisation pour un focus.C'est ce que nous avons fait, en introduisant les notions de normalisation par le maximum et de normalisation par la somme. Ensuite nous avons étudié les propriétés que nous désirons pour une mesure de comparaison entre focus, ce qui nous a permis de définir la pertinence relative d'un focus par rapport à un autre, mesure qui n'est pas une distance.*

ABSTRACT. *Semantic indexing of documents using ontologies is a growing field, despite difficulties with automatic or semiautomatic indexing, not to mention manual indexing. Nowadays we have semantic characterizations of textual or non textual documents based on ontologies.Hence, we developped an entity, the* focus*, that represents a document or a query with wheightings on concepts of an ontology in order to indicate the importance of each concept within the document. Nevertheless, that solution needs for precise definition of focus normalization. Introducing notions of normalization by the maximum and normalization by the sum enabled us to accomplish this. Thus, we study the properties we want for a similarity measure between focusses, which allows us to define the relative relevance of a focus in comparison with another one.*

MOTS-CLÉS : *ontologie, qualification sémantique, recherche d'information distribuée.*

KEYWORDS: *ontology, semantic characterization, distributed information retrieval.*

1. Introduction

La quantité d'information sous format électronique a augmenté de manière critique depuis quelques années. On[1] estime cette augmentation à plus de 30% par an entre 1999 et 2002 en ce qui concerne les nouvelles informations, et le développement de nouveaux systèmes (téléphones mobiles, pair-à-pair, appareils photos numériques, etc.) accélère le mouvement : la messagerie instantanée génère cinq milliards de messages par jour (750 Giga Octets), ou 274 Tera Octets par an, les appels téléphoniques dans le monde mettent en jeu 17,3 Exa Octets de nouvelles informations si on les stockait sous un format électronique, etc.

Aider un utilisateur à trouver l'information qu'il cherche dans ce contexte devient donc de plus en plus difficile. D'autant plus que la recherche d'information (RI) fait face à deux autres problèmes : le problème du vocabulaire et le problème de l'hétérogénéité des données.

Le premier est bien connu des utilisateurs de moteurs de recherche. En effet, lorsque nous utilisons un de ces derniers pour effectuer une recherche sur le web, nous faisons face au problème du choix des mots-clés et à celui du filtrage des réponses. Ces problèmes reposent sur la capacité des mots des langages humains à générer polysémie (plusieurs sens pour un mot) et synonymie (plusieurs mots ayant le même sens). [FUR 87] par exemple montrent les difficultés de consensus dans le choix des termes pour les indexations et pour les recherches. La probabilité que le même terme soit choisi est inférieure à 20%. Et même si on utilise un thésaurus contraint, avec une liste de mots acceptés (par exemple, pour des formulaires de saisie avec des codes ou des intitulés pré-définis), la probabilité ne dépasse pas 70%.

Le second problème concerne les différences de langage (français, anglais, etc.), de format (texte, vidéo, image, etc.), de type (.tex, .html, .pdf, etc.) des résultats d'une requête. En effet, une réponse pertinente à une interrogation quelconque traverse ces différences et l'utilisation de recherche syntaxique oblige à des traductions, des descriptions, entre les différents langages de représentation des données.

Nous avons décidé de passer à un niveau sémantique pour dépasser ces difficultés classiques de la recherche d'information. Dans ce but, nous avons mis en place une entité nouvelle, le *focus*, c'est-à-dire une pondération des concepts d'une ontologie. Il est facilement échangeable (liste de pondérations) et il permet des descriptions de haut niveau ; il s'agit donc de caractérisations sémantiques « universelles », utilisables pour des textes, images, vidéos, de différentes langues, de différents formats, etc.

La suite de cet article est la suivante : dans une première partie nous définissons le focus (section 2). Puis nous mettons en place des normalisations de celui-ci dans le but de pouvoir le comparer (sections 3 et 4). Enfin nous étudions les résultats de notre approche (section 5).

1. Source : http ://www.sims.berkeley.edu/research/projects/how-much-info-2003/execsum.htm

2. Le focus

2.1. *Ontologies*

L'objectif des ontologies est de fournir des représentations des connaissances pour les systèmes informatiques. Elles sont composées de concepts (ou types de concepts) et de relations (ou types de relations) entre ces concepts. Parfois des axiomes sont mêmes définis pour indiquer quelques règles ou propriétés générales sur les entités et les relations.

L'ontologie que nous trouvons le plus fréquemment dans les applications du domaine est WordNet[2], bien qu'elle ne soit à proprement parler qu'un thésaurus (les « concepts » sont des listes synonymes, la grande majorité des liens représentés sont des liens de subsomption, etc.). C'est donc celle que nous allons utiliser, nous limitant par conséquent à des connaissances terminologiques ou conceptuelles, et à la relation hiérarchique *IS-A*.

2.2. *Le focus en général*

Un focus est une application définie sur une ontologie Ω. Soit c_Ω l'ensemble de ses concepts :

$$\forall c_i \in c_\Omega,\ \vec{f}[c_i] \in [\min, \max] \tag{1}$$

$\min$ et $\max$ sont les bornes des valeurs réelles que peut prendre chaque concept de l'ontologie. Nous avons choisi une notation vectorielle : nous notons $\vec{f}[i]$ la valeur du i-ème concept du focus $\vec{f}$. Pour des raisons de simplicité, nous supposons que $\min$ vaut 0.

Les systèmes utilisant des caractérisations sémantiques que nous avons trouvés dans la littérature, par exemple [HAL 03], ont une gestion centralisée de celles-ci, c'est-à-dire qu'un agent comparateur reçoit caractérisations de documents et les traite lui-même, c'est lui qui regarde si les concepts des requêtes et des documents sont proches, suivant une mesure de proximité sémantique.

Cette approche ne nous convient pas. En effet, les agents fournisseurs de contenus ne sont pas autonomes, ce ne sont pas eux qui choisissent si *chat persan* doit être proche ou non de *chat*. Le comparateur peut en effet décider que le concept c_i présent dans un document du fournisseur F_k est proche du concept c_j de la requête r, mais il est tout à fait possible que pour F_k ces concepts ne soient pas proches.

Selon nous, les fournisseurs doivent assumer leurs caractérisations. Cela suppose néanmoins que tous les concepts pertinents pour un fournisseur soient pondérés. Il est donc nécessaire de mettre en place une mesure de propagation des pondérations des concepts dans une ontologie. C'est ce que nous avons fait en [VEN 04].

2. http ://wordnet.princeton.edu/

Nous avons donc maintenant des caractérisations sémantiques de documents et de requêtes, les focus. Il s'agit de pondérations de concepts d'une ontologie. Mais toutes sortes de pondérations peuvent être présentes dans les focus : un fournisseur peut avoir mis 10000 sur un concept en voulant signifier une grande importance sur ce concept ; et un autre fournisseur peut estimer que 50 est une valeur maximum. Comment réussir à comparer dans ce cas les focus ?

3. Normalisation de focus

3.1. *Normalisation par la somme*

Pour comprendre cette normalisation, il faut penser à une table de jeu à la roulette : il y a des jetons (des pondérations possibles) et des valeurs (des concepts). La somme que nous pouvons miser est limitée et ainsi, mettre un nombre important sur une valeur a un sens, cela signifie que nous lui donnons une grande importance. Normaliser consiste ici à mettre la même somme de pondérations sur les focus, ce qui permet d'indiquer l'intérêt relatif de chaque concept dans le focus.

Définition 1 *Un focus $\vec{f}$ est* normalisé à k par la somme *si et seulement si :* $\sum_{i\in[1..|\vec{f}|]} \vec{f}[i] = k$ *avec k une constante.*

Le processus de normalisation est des plus simples. Soit $\vec{f}$ un focus. Sa forme normale par la somme à k, notée $\vec{f}^{\Sigma=k}$, ou plus simplement $\vec{f}^{\Sigma}$ lorsqu'il n'y a pas d'ambiguïté sur k, s'obtient :

$\forall i \in [1..|\vec{f}|]$:

$$(\vec{f})^{\Sigma}[i] = \begin{cases} \frac{k}{\sum_i \vec{f}[i]}\vec{f}[i] & \text{si } \vec{f} \neq \vec{0} \\ 0 & sinon \end{cases}$$

3.2. *Normalisation par le maximum*

Cette fois-ci, l'intuition est l'échelle de valeur, avec un maximum. Si les valeurs sur tous les focus ont une borne maximum, alors il est possible de remarquer les différences relatives d'importance mise sur le même concept dans différents focus.

Définition 2 *Un focus est* normalisé à k par le maximum *si et seulement si :* $\max_{i\in[1..|\vec{f}|]} \vec{f}[i] = k$ *avec k une constante.*

Soit $\vec{f}$ un focus. Sa forme normale par le maximum à k, notée $\vec{f}^{\max=k}$, ou plus simplement $\vec{f}^{\max}$ lorsqu'il n'y a pas d'ambiguïté sur k, s'obtient de la même façon que précédemment.

4. Pertinence d'un focus par rapport à un autre

Dans le modèle vectoriel des documents [BER 99], modèle qui ressemble à notre approche, la mesure de similarité la plus utilisée est le cosinus. Nous pouvons la réécrire ainsi avec les focus :

$$cosinus(\vec{f_1}, \vec{f_2}) = \frac{\sum_i (\vec{f_1}[i].\vec{f_2}[i])}{\sum_i (\vec{f_1}[i].\vec{f_1}[i]).\sum_i (\vec{f_2}[i].\vec{f_2}[i])}$$

Cette mesure satisfait les propriétés d'une distance : minimalité, symétrie, inégalité triangulaire. Néanmoins, lors d'une étude un peu approfondie sur le raisonnement analogique et métaphorique en psychologie, nous sommes arrivé à des constatations qui mettent en cause ce choix de la littérature. En effet, [TVE 77] montre que les mesures de similarité telles que les humains les utilisent ne sont pas des distances. En effet, s'il est possible que nous disions qu'un fils ressemble à son père, l'inverse est beaucoup plus rare. De même si un triangle rouge peut selon nous être proche d'un carré rouge (parce qu'ils partagent la même couleur), et que ce dernier peut être proche d'un carré jaune (même forme), on ne peut pas dire que la proximité entre le triangle rouge et le carré jaune soit inférieure aux deux dernières. C'est pourquoi nous avons mis en place une mesure de pertinence relative d'un focus par rapport à un autre, mesure qui ne respecte ni la symétrie, ni l'inégalité triangulaire.

$$PR(\vec{f_1}, \vec{f_2}) = \sum_i (comm(\vec{f_1}[i], \vec{f_2}[i]) - diff(\vec{f_1}[i], \vec{f_2}[i]))$$

avec :

$$comm(\vec{f_1}[i], \vec{f_2}[i]) = \min(\vec{f_1}^{max}[i], \vec{f_2}^{max}[i]) \times \vec{f_2}^{\Sigma}[i]$$

$$diff(\vec{f_1}[i], \vec{f_2}[i]) = \begin{cases} (\vec{f_1}^{max}[i] - \vec{f_2}^{max}[i]) \times \vec{f_1}^{\Sigma}[i] \\ \text{si } \vec{f_1}^{max}[i] > \vec{f_2}^{max}[i] \\ 0 \qquad\qquad\qquad\qquad\qquad \text{sinon} \end{cases}$$

En ce qui concerne $comm(\vec{f_1}[i], \vec{f_2}[i])$ il s'agit de prendre en considération la partie commune aux deux focus (par le $\min$), mais en le pondérant par l'intérêt relatif ($\vec{f}^{\Sigma}$) du focus cible. Ainsi, nous obtenons une partie commune entre les deux focus, mais suivant l'intérêt qu'y porte le focus cible. Pour $diff(\vec{f_1}[i], \vec{f_2}[i])$ nous nous intéressons aux différences entre le focus pivot et le focus cible, dans ce sens-là seulement. Nous pondérons cette différence par l'intérêt relatif que le focus pivot porte aux différents concepts.

5. Résultats

Il est tout d'abord à noter que la complexité du calcul de similarité que nous avons défini est en $O(2n)$ (deux parcours de vecteurs). Or, les approches classiques utilisant

des pondérations de concepts dans un contexte distribué le font en $O(n^2)$. Ou bien, il leur faut mettre en place des solutions centralisées, avec des matrices récapitulant tous les documents. Pour nos tests, nous utilisons un Pentium III à 1 Ghz, avec 256 Mo de mémoire vive. L'ontologie commune à tous les focus comprend 389 concepts et, pour chaque focus, nous avons choisi au hasard un certain nombre de concepts (une cinquantaine en moyenne). Le tableau 1 récapitule les résultats. Nous remarquons que

	100	100000	500000
moyenne	27.8	1240.5	5263.6

Tableau 1. *Tests de comparaisons de focus : un focus (requête par exemple) est comparé à un certain nombre d'autres focus (documents). Les temps sont en millisecondes, les colonnes correspondent à 100, 100000 et 500000 focus.*

les comparaisons sont rapides : il faut en moyenne 5 secondes pour comparer 500000 focus de documents à un focus de requête, et la progression est linéaire.

6. Conclusion

Face aux problèmes d'hétérogénéité des langages de représentation des données et à l'ambiguïté de nos langues naturelles, nous avons mis en place des caractérisations sémantiques de documents et de requêtes. Ces caractérisations, les focus, sont normalisées et peuvent donc être comparées.

Nous avons aussi défini une mesure de pertinence relative qui ne soit pas une distance, conformément aux constatations de [TVE 77] en psychologie. Ainsi nous pouvons faire des comparaisons entre focus qui soient fidèles aux modèles humains de comparaison. Cette approche est enfin d'une complexité de calcul faible.

7. Bibliographie

[BER 99] BERRY M. W., DRMAC Z., R.JESSUP E., « Matrices, Vector Spaces, and Information Retrieval », *SIAM Rev.*, vol. 41, n° 2, 1999, p. 335–362, Society for Industrial and Applied Mathematics.

[FUR 87] FURNAS G., LANDAUER T., GOMEZ L., DUMAIS S., « The vocabulary problem in human-system communication », *Communications of the Association for Computing Machinery*, vol. 30, 1987, p. 964–971.

[HAL 03] HALKIDI M., NGUYEN B., VARLAMIS I., VAZIRGIANNIS M., « THESUS : Organizing Web document collections based on link semantics », *The VLDB Journal*, vol. 12, n° 4, 2003, p. 320–332, Springer-Verlag New York, Inc.

[TVE 77] TVERSKY A., « Features of Similarity », *Psychological Review*, vol. 84, n° 4, 1977, p. 327–352.

[VEN 04] VENTRESQUE A., « Focus et ontologie pour la recherche d'information », Mémoire de DEA d'informatique, Université de Nantes, France, 2004.