Utilisation d'une architecture ontologique dans un système d'aide à la recherche d'information en ligne(SARIOnto)

Hajer baazaoui zghal* — Rania Soussi* — Nesrine Ben Mustapha *— Marie-aude Aufaure** — Henda Hajjami Ben Ghézala*

*Laboratoire RIADI ENSI Campus Universitaire de la Manouba 2010

** Supélec Plateau du Moulon 3, rue Joliot Curie 91 192 Gif sur Yvette Cedex

{nesrine.benmustapha, , hajer.baazaouizghal, henda.benghezala}@riadi.rnu.tn

Marie-Aude.Aufaure@supelec.fr

RÉSUMÉ. La croissance très importante des informations disponibles sur Internet nécessite des outils de recherche de plus en plus performants permettant de discerner efficacement les informations intéressantes parmi des centaines voire des milliers de documents. Seulement, la qualité des résultats fournis par les moteurs de recherche traditionnels n'est pas toujours pertinente surtout quand il s'agit de composer plus d'une requête. Ceci est dû aux ambiguïtés linguistiques et aux concepts abstraits qui ne sont pas bien traités. L'utilisation de la sémantique et plus précisément des ontologies présente des atouts importants. L'objectif de cet article est de montrer l'apport des ontologies dans la recherche d'information en ligne. Ainsi, un système d'aide à la recherche d'information en ligne basé sur les ontologies est proposé. Ce système est composé de deux ontologies : une ontologie de domaine et une ontologie de services ainsi que WordNet, pour représenter les concepts ainsi que les services de domaine. La contribution de ces ontologies pour améliorer la recherche d'information en ligne est montrée par l'implémentation et l'expérimentation du système proposé.

ABSTRACT. The huge number of available documents on the Web makes finding relevant ones challenging. Thus, searching for information becomes more and more complex because of the growing volume of data and of its lack of structure. The quality of results that traditional full-text search engines provide is still not optimal for many types of user queries. Especially, the ambiguities of natural languages and abstract concepts are handled inadequately by full-text search engines. Ontologies provide a solution to these problems. For adopting this solution, a composed architecture of several ontologies is proposed to represent concepts as well as services of domain. In this paper, we expose the contribution of these ontologies in information retrieval and we propose a new retrieval system based on ontologies. Experimentation in the domain of the tourism is presented, and the gotten results are compared to other systems.

MOTS-CLÉS : recherche d'information en ligne, construction d'ontologies

KEYWORDS: on-line Information retrieval, ontologies construction

1. Introduction

Depuis l'avènement de l'Internet dans les années 90, le nombre croissant de documents disponibles rend indispensable l'utilisation d'outils adaptés permettant d'assister et d'aider les utilisateurs au cours de leurs recherches d'information en ligne. Par ailleurs, les techniques généralement employées par les moteurs de recherche reposent sur des méthodes statistiques et des traitements syntaxiques qui ne permettent pas de traiter la sémantique contenue dans la requête de l'utilisateur ainsi que dans les documents. D'autres problèmes, tels que celui lié au vocabulaire et à l'hétérogénéité des données, le choix des mots clés et du filtrage peuvent être rencontrés. Ces problèmes reposent sur la capacité des mots des langages humains à générer la polysémie et la synonymie. Aider un utilisateur à trouver l'information qu'il cherche dans ce contexte devient donc une tâche de plus en plus difficile. D'où le recourt à la prise en compte du niveau sémantique qui permet de pallier ces difficultés de la recherche d'information. Des approches ont été proposées pour permettre l'extraction de la sémantique et mieux répondre aux requêtes émises. Par contre, la plupart de ces techniques ont été conçues pour s'appliquer au Web en entier et non pas à un domaine particulier. Une piste intéressante consiste à utiliser une ontologie non seulement pour représenter un domaine spécifique mais aussi pour le filtrage et la classification des documents retournés à l'utilisateur par sous thèmes. Dans cet article, nous exploitons cette piste et exposons dans la première section une architecture ontologique pour la représentation sémantique des documents sur le Web. Ensuite, nous présentons notre système de recherche d'information en ligne basé sur deux ontologies de l'architecture exposée. L'évaluation de notre système est présentée à la fin de cet article.

2. Architecture ontologique pour la recherche d'information sur le Web.

Dans le Web actuel, pour spécifier les connaissances d'un domaine donné, nous distinguons trois types de sémantiques [2]: la sémantique du domaine, la sémantique des services du domaine et la sémantique de la structure des sites Web. Le Web actuel présente une absence des relations explicites entre ces trois sémantiques. Pour résoudre cette première problématique, une architecture qui comprend trois ontologies est proposée l'ontologie de domaine, l'ontologie de structure des sites Web et l'ontologie des services de domaine [1]. À ce niveau, nous avons supposé que la spécification des connaissances d'un domaine particulier dans le Web ne peut pas être fournie seulement par l'ontologie de domaine, du fait qu'une même ontologie ne peut pas spécifier les trois sémantiques et les relations qui existent entre elles en même temps. Ainsi, l'architecture proposée, est construite autour de trois ontologies interdépendantes et est dotée d'un processus incrémental d'apprentissage d'ontologies. L'ontologie de domaine permet la définition

d'un domaine donné en termes de concepts, de relations et d'axiomes. L'ontologie de services du domaine présente une autre vue du domaine en mettant en valeur ses différents acteurs, services, sous-services, activités, tâches et processus de déroulement de chacun des services. L'ontologie de structure des sites Web spécifie les structures communes des sites Web relatives aux services du domaine. Dans la section suivante, nous détaillons notre système de recherche d'information (RI) en ligne basé sur les ontologies: une ontologie de domaine et une ontologie de services.

3. Ontologie et recherche d'information

Dans l'objectif de fournir à l'utilisateur un accès facile et pertinent à l'information à laquelle il s'intéresse, certains des travaux actuels dans la RI tentent d'améliorer le procédé de récupération avec l'aide des ontologies. Les expériences effectuées en utilisant WordNet [5] avec une stratégie intelligente de recherche ont prouvé qu'un gain significatif est possible sur le TREC (Text Retrieval conference) de l'ensemble des données [3] [10]. Elles peuvent aider l'utilisateur à détecter ses besoins et trouver les mots-clés appropriés qui utilisent les concepts existants dans l'ontologie et leur description. La connaissance représentée par les ontologies peut être utilisée à différents niveaux dans le processus de RI. Elle peut aider à l'indexation des documents, alors appelée indexation sémantique. Les ontologies peuvent également aider à la formulation du besoin de l'utilisateur et à l'accès aux documents. Enfin l'ontologie peut être utilisée dans le modèle lui-même pour réaliser l'appariement entre le besoin et les documents. Afin de profiter des apports des ontologies dans la recherche d'information, nous proposons un système d'aide à la recherche d'information en ligne basé sur les ontologies. Ce système repose sur deux ontologies : une ontologie de domaine et une ontologie de services ainsi que WordNet. Dans ce qui suit, nous détaillons la construction des ontologies, l'architecture de SARIOnto et l'apport de ces ontologies.

3.1. Construction des ontologies

L'ontologie de domaine est construite semi-automatiquement en utilisant notre technique d'apprentissage d'ontologies développée dans OntoCoSemWeb [1]. Cette ontologie est la base de notre système SARIOnto, et contient la plupart des concepts du domaine et leurs propriétés. Chaque domaine est caractérisé par une liste de services, activités et tâches. Ces services sont liés aux concepts existant dans l'ontologie du domaine. Chaque concept de l'ontologie du domaine est le sujet d'un ou plusieurs services, activités ou tâches. Par exemple, le service «Lodging_hotel » est associé au concept « hotel ». Cette relation donne la possibilité d'améliorer la recherche et aider les utilisateurs à mieux exprimer leurs besoins. La construction de cette ontologie des services,

ainsi que la mise en correspondance avec l'ontologie de domaine, sont pour le moment réalisées manuellement.

3.2. Architecture de SARIOnto

Le système proposé est constitué de trois principaux modules:

- Un module de traitement de la requête qui permet d'enrichir la requête initiale de l'utilisateur en se basant sur les concepts et relations de l'ontologie de domaine et sur WordNet.
- Un module de traitement des documents qui traite le résultat obtenu suite à la recherche basée sur la requête enrichie.
- Un module de classification par service qui assure la classification des documents par service et guide l'utilisateur à construire une deuxième requête en utilisant les services choisis.

3.3 Utilisation des ontologies dans SARIOnto

Dans cette partie nous détaillons le fonctionnement des modules de SARIOnto.

3.3.1. Enrichissement de la requête

Afin d'obtenir une requête plus pertinente qui met en relief le domaine de recherche, la requête émise par l'utilisateur est enrichie sur la base des concepts et des relations de l'ontologie du domaine et les relations sémantiques de WordNet. SARIOnto permet à l'utilisateur de s'appuyer sur l'ontologie de domaine pour choisir les concepts qu'il veut ajouter à sa requête. La requête est traitée en plusieurs étapes, qui sont :

Analyse morphologique : Elle permet la reconnaissance des différentes formes de mots à partir d'un lexique (dictionnaire, thésaurus). La lemmatisation permet la transformation d'une forme conjuguée ou fléchie à sa forme canonique ou lemme (exemple : constitutionnel constitution). Dans notre approche la lemmatisation de la requête est assurée par TreeTagger [7].

Analyse sémantique : Après la lemmatisation de la requête, les lemmes des termes sont filtrés. Les mots vides sont éliminés, nous ne traitons que les verbes (verbe non vide : différent de find, do, get…) et les noms pour détecter par la suite les concepts correspondants existant dans l'ontologie de domaine. Pour poursuivre l'analyse sémantique, nous utilisons WordNet, qui fournit les différents sens associés aux mots de la langue anglaise (par exemple, le mot room est associé à 5 sens). Il s'est ainsi, avéré intéressant de permettre aux utilisateurs de sélectionner les sens des mots clés. Puis, l'algorithme suivant est appliqué aux termes lemmatisés :

- Si le terme est un verbe, nous conservons sa forme lemmatisée. Par exemple : le verbe « traveled » a pour lemme « travel », c'est ce lemme qui est sauvegardé.

- Si le terme est un nom : ses synonymes, hyponymes et hypéronymes sont récupérés à partir de WordNet, en utilisant le sens choisi par l'utilisateur. Deux sous cas se présentent, (1) si le terme ou l'un de ses composants existe dans l'ontologie alors le terme sera remplacé par ce concept, ses sous concepts et les propriétés qui lui sont relatifs, (2) si ni le terme ni ses composants ne sont présents dans l'ontologie alors ce terme est éliminé de la requête.

Ainsi, nous obtenons une nouvelle requête construite à partir des concepts et des relations extraites à partir de l'ontologie de domaine. Par exemple, la requête initiale envoyée par l'utilisateur est: "find hotel ". Cette requête est lemmatisée avec TreeTagger. La forme lemmatisée rendue est: find VV find, hotel NN hotel(terme, type grammatical, et lemme). Le mot "hotel" a un sens unique dans Wordnet et "find" est un verbe vide, il est ainsi enlevé de la requête. L'analyse sémantique retourne une requête enrichie qui est "Hotel Suite Room Address Name Price Star".

3.3.2. Processus de recherche

Nous utilisons l'ontologie pour généraliser des mots à des concepts comme il sera expliqué dans les étapes ultérieures. Après son enrichissement, la requête est soumise à un moteur de recherche (google API). Les documents résultants sont traités comme suit:

Analyse sémantique : Chaque document, se trouvant dans la liste (qui n'est pas dans le format pdf ou doc), est téléchargé à partir de son URL sous la forme d'un document HTML puis analysé avec DOM [9]. En récupérant le texte et en l'analysant morphologiquement avec TreeTagger, nous pouvons extraire les formes de base des mots pour pallier les problèmes des variations morphologiques et détecter les concepts qui existent dans l'ontologie de domaine et dans la requête.

Filtrage avec le modèle vectoriel : Pour récupérer les documents les plus pertinents, Nous avons utilisé le modèle vectoriel de Salton [6]. Mais, pour que ce modèle soit plus adapté à notre approche, nous substituons les concepts aux termes. Plus précisément, dans la formule de similarité nous calculons les poids des concepts au lieu de ceux des termes. Un concept est représenté par l'ensemble de ses synonymes, hypéronymes et hyponymes. Dans ce modèle chaque document est représenté par un vecteur : Soit ti, un concept de la requête Q, Dj = (d1j, d2j, d3j, …, dNj), avec dij : poids du concept ti dans le document Dj, et N est le nombre de concepts qui forment la requête. Et chaque requête est représentée par un vecteur : Q = (q1, q2, q3, …, qN), qi : poids du concept ti dans la requête Q.

Dans notre approche, les poids des termes dans la requête valent toujours initialement 1 (la requête est reformulée avec les concepts et relations de l'ontologie, et sans répétition du même terme). Plus explicitement, si notre requête reformulée est par exemple « hotel room », alors Q= (1,1). Pour les documents, nous utilisons la formule de poids TF*IDF

normalisée [8] pour donner une chance égale à tous les documents et ne pas favoriser les plus longs. La mesure de similarité entre la requête et les documents sera calculée avec la formule de cosinus [7]. Après avoir calculé la similarité pour tous les documents, nous obtenons une deuxième liste triée par ordre croissant suivant leurs similarités par rapport à la requête. Seuls les documents qui ont une similarité non nulle sont affichés à l'utilisateur.

3.3.3. Classification des documents et reformulation de la requête

La classification permet d'affecter un document à une des classes ou sous-classes d'un plan de classement ou table de classification, c'est-à-dire dans un des domaines de la connaissance. Un résultat classifié par catégorie permet de faciliter la récupération de l'information et même de détecter d'autres besoins. Lorsque le résultat est traité et affiché à l'utilisateur, il peut faire l'objet d'une classification par service fournie par le module de classification. La requête reformulée par le module de traitement de la requête contient un ensemble de concepts qui appartiennent à l'ontologie du domaine. Ces derniers sont en liaison avec un ensemble de services de l'ontologie de services. En utilisant cette liaison, nous pouvons extraire, à partir des concepts ajoutés aux requêtes, les services qui lui sont relatifs dans l'ontologie des services. Le modèle vectoriel est de nouveau utilisé, un service est représenté par un vecteur : Servi= (c1, c2,….cN) avec N le nombre de concepts relatifs à un service. Pour chaque document retenu, nous calculons sa similarité avec les services en utilisant la formule du cosinus, puis il est affecté au service avec lequel il est le plus similaire. Ainsi, les URL sont affichées par service. Dans l'exemple précédent, la requête initiale était "find hotel". Les services, activités et tâches détectées à partir des concepts dans l'ontologie du service sont: Le service: Lodging_Hotel. Les activités: Reservation, Hotel_Search, Modify_Reservation, Disponibility_Verification. Les tâches: Verify_RoomNumber, Verify_RoomType, Search_HotelName, Search_HotelAdress.

Chaque document ayant une valeur de similarité supérieure à 0,3(ce sont les documents les plus similaires à la requête) est attribué à un service, une activité et une tâche. L'utilisateur peut alors choisir les services correspondants à ce qu'il recherche. Le module de traitement de la requête capture les services choisis et utilise la liaison entre l'ontologie de services et celle de domaine pour formuler une deuxième requête avec les nouveaux concepts et les relations détectés. La nouvelle requête est envoyée au moteur de recherche et le système réitère le processus déjà décrit pour afficher un nouveau résultat plus raffiné. Dans notre exemple, si l'utilisateur choisit la tâche "Search_Hotel_Name", la nouvelle requête sera "Hotel Name" parce que le concept "Hotel" et la propriété "Has_Name" sont liés à cette tâche. Une évaluation a été menée et est décrite dans la section suivante.

4. Evaluation de SARIOnto

SARIOnto permet de fournir à l'utilisateur un service en ligne. Il utilise l'API Jena[1]pour manipuler les ontologies et Google API pour effectuer les recherches à travers le web. Dans le but d'évaluer le système proposé, nous utilisons les mesures de précisions et rappel classique. En revanche, évaluer des systèmes de recherche basés sur la sémantique est une tâche complexe vu que les mesures classiques de précision/rappel s'avèrent difficile à être calculée de manière automatique. On a choisi un protocole d'évaluation centré utilisateurs. En effet, nous avons effectué un test basé sur 15 requêtes dans le domaine du tourisme et évalué par 10 utilisateurs.

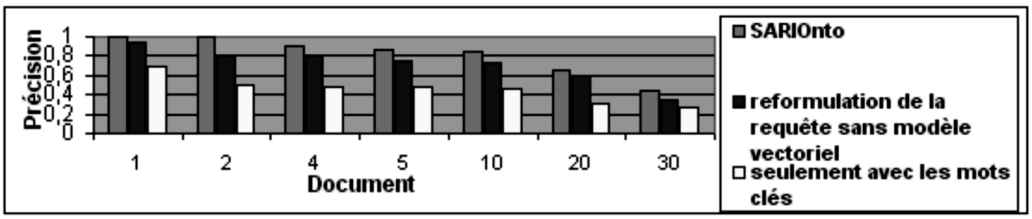

Figure 1. *Comparaison de SARIOnto à d'autres systèmes de recherche*

Puis, nous avons comparé ces mesures avec les mesures obtenues en utilisant deux autres systèmes; à savoir le premier, Lucene[2] est un système de recherche d'information traditionnel basé sur la recherche par les mots-clés. Le deuxième est basé sur la reformulation de la requête, mais sans faire appel au modèle vectoriel. La figure 1 présente les résultats en termes de précision. Notre système est meilleur que les deux autres systèmes en terme de précision, par exemple pour 30 documents retournés, SARIOnto offre une amélioration de 20% de précision par rapport à lucene et 10% par rapport à l'autre système.

5. Conclusion et perspectives

La qualité des résultats fournis par les moteurs de recherche traditionnels n'étant pas toujours pertinente pour plusieurs requêtes ceci a conduit les chercheurs à trouver des solutions reposant sur les ontologies. C'est dans ce cadre que s'inscrit le travail présenté par cet article, qui propose un système de recherche d'information en ligne basé sur cette architecture, permettant d'améliorer la pertinence des résultats présentés à l'utilisateur. L'expérimentation et l'évaluation menées montrent une amélioration du taux de précision comparé à d'autres systèmes. Notre proposition permet de détecter les services relatifs à un

[1] http://jena.sourceforge.net/
[2] http://lucene.apache.org

domaine précis, et de retourner un ensemble de documents classés. Nos travaux en cours cherchent à permettre à l'utilisateur d'importer une ontologie de domaine pour effectuer ses recherches. Une autre perspective concerne l'enrichissement de notre système par une ontologie floue pour améliorer son efficacité. Ce travail permet aussi de fournir des documents pré classés et filtrés pour améliorer la construction des composants ontologiques [4].

6. Bibliographie

[1] Baazaoui Zghal, H., Aufaure, MA. and Ben Mustapha, N. (2007). Extraction of Ontologies from Web Pages: conceptual modeling and tourism, *Journal of internet Technologies,* Volume 8 No. 4 (2007), octobre 2007, ISSN 1607-9264.

[2] Baazaoui Zghal, H., Aufaure, MA. and Ben Mustapha, N. (2007). A Model-Driven approach of ontological components for on-line semantic web information retrieval, *Journal on Web Engineering*, Special Issue on Engineering the Semantic Web, Rinton Press, December 2007, Volume 6 No. 4, 309-336.

[3] Baziz M. (2004) Towards a Semantic Representation of Documents by Ontology-Document Mapping. *The Eleventh International Conference on Artificial Intelligence*(AIMSA 2004)

[4] Ben Mustapha, N., Baazaoui Zghal, H. and Aufaure, MA. (2007). A Prototype for knowledge extraction from semantic web based on ontological components construction, Internet technology, *Web information System and Technologies* (WEBIST07).

[5] Miller, G.A. (1995) WordNet: A Lexical Database for English. *Communications of the ACM*, 11, 39-41.

[6] Salton G., MacGill M.J. (1983). *Introduction to modern information retrieval* , McGraw Hill International Book Company, ISBN 0-07-Y66526-5.

[7] Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. IMS-CL, Institut Für maschinelle Sprachverarbeitung, Universität Stuttgart, Germany.

[8] Singhal, A., Salton, G., Mitra, M. et Buckley, C. (1995). Document length normalization. *Information Processing and Management*, 32(5), pages 619–633.

[9] Stenback, J. , Le Hors, A., Le Hégaret, P. (2003). Document Object Model (DOM) Level 2 HTML Specification, http://www.w3.org/TR/, 9.

[10] Voorhees, E. M., Gupta, N. K. et Johnson-Laird, B. (1994). The collection fusion problem. *Proceedings of the 3rd Text REtrieval Conference* (TREC-3). p. 95-104.