

Interprétation sémantique des requêtes utilisateurs par usage des ontologies

Hanéne DGHIM, Habib SMEI, Abdelmajid BEN HAMADOU

Laboratoire de recherche LARIM Institut Supérieur d'Informatique et de Multimédias de Sfax Tunisie
Dghim_hanene@yahoo.fr, habib.smei@isetsf.rnu.tn , abdelmajid.benhamadou@isimsf.rnu.tn

Résumé :

Le développement considérable qu'a connu Internet ces dernières années notamment avec l'apparition du World Wide Web, a conduit à une croissance quasi exponentielle du nombre d'utilisateurs et aussi du nombre de documents mis à disposition.

Les systèmes de recherche d'informations (SRI) sont confrontés à un nouveau défi relatif à la pertinence des documents obtenus en réponse à une requête utilisateur.

Dans cet article nous proposons d'exploiter les ontologies pour l'interprétation sémantique des requêtes utilisateurs pour améliorer la recherche d'informations dans le cadre de l'élaboration d'un SYstème de Fouille Automatique et d'eXtraction des documents pédagogiques (SYFAX).

Le processus d'interprétation sémantique commence par dissocier le type de document recherché de son contenu en exploitant une ontologie du domaine pour les types des documents. .

Une fois le type de ressources recherchés est connu, nous pouvons utiliser l'ensemble des mots clés utilisés dans la requête pour la recherche dans l'entrepôt SYFAX ,des documents répandant à ces mots clés. Nous nous basons dans cette phase aussi sur une ontologie du domaine qui nous permette d'enrichir le vocabulaire utilisé dans la requête pour étendre le champ de recherche.

Mots-clés : document pédagogique, Système de Recherche d'Information, ontologie, interprétation sémantique des requêtes, expansion des requêtes.

The considerable development that Internet knew these last years in particular with the appearance of the World Wide Web, led to a quasi exponential growth of the number of users and also the number of documents available.

Information Retrieval Systems (IRS) are confronted to a new challenge relating to the relevance of the documents obtained in result of a user query.

In this article we propose to exploit ontologies for the semantic interpretation of users queries to improve search of information within the setting of an automatic System of search and extraction of pedagogical documents (SYFAX). The process of semantic interpretation starts by dissociating the kind of required documents of its contents by exploiting an ontology for the documents types. Once the type of resources required is known, we can used the whole of the key words used in the query for search in SYFAX warehouse, of the documents that matching with these key words. In this phase we also use a domain ontology which enables us to enrich the vocabulary used in the request to extend the field of search.

Key words: pedagogical document, Information Retrieval System, ontology, semantic interpretation of queries, queries expansion.

1. Introduction

Le développement considérable qu'a connu Internet ces dernières années notamment avec l'apparition du World Wide Web, a conduit à une croissance quasi exponentielle du nombre d'utilisateurs et aussi du nombre de documents mis à disposition : Le nombre d'utilisateurs est aujourd'hui évalué à des centaines de millions et le nombre de pages Web accessibles a augmenté de 320 millions en 1997 à plus de 4 milliards en 2005.

Les systèmes de recherche d'informations (SRI) sont confrontés à un nouveau défi relatif à la pertinence des documents obtenus en réponse à une requête utilisateur.

En effet, les SRI utilisent le vocabulaire de la langue comme lien de correspondance entre la requête et le document (correspondance requête/contenu). Néanmoins, cette stratégie basée sur la sémantique de la langue, se révèle dans la pratique insuffisante comme le montre l'augmentation du bruit lors d'une recherche.

En saisissant une requête composée par exemple des mots « cours » et « algorithme », l'utilisateur reçoit en réponse à sa requête, tous les documents dont les mots clés contiennent les termes « cours » et/ou « algorithme », alors que le mot « cours » désigne ici le type de documents recherchés par l'utilisateur et non pas un mot clé. Ceci vient du fait que les SRI classiques n'effectuent aucune interprétation sémantique des termes utilisés dans la requête.

L'émergence du « Web sémantique » aux années 90, a été dont le but de remédier aux insuffisances des SRI classiques. un intérêt particulier a été donné à la sémantique des requêtes utilisateurs afin d'affiner d'avantage le processus de recherche. Plusieurs sujets ont abordé l'analyse des requêtes utilisateurs et l'une des solutions proposées était l'application des ontologies pour l'expansion des requêtes utilisateurs. Mais ceci n'a pas résolu le problème puisque le rôle des ontologies était d'enrichir une requête utilisateur par d'autres termes du même domaine sémantique que les mots de la requête initiale, sans aucune interprétation sémantique possible de la requête.

L'interprétation sémantique des requêtes utilisateurs par l'usage des ontologies fait l'objet de notre travail qui s'inscrit dans le cadre de l'élaboration d'un Système de fouille automatique et d'extraction des documents pédagogiques (SYFAX) [SFX1] [SFX2], conçu dont le but de fournir à une communauté (i.e., universitaire) un portail facilitant le partage des documents pertinents aux membres de cette communauté.

Cet article est organisé comme suit :

Dans la section 2 nous décrivons quelques travaux qui ont utilisé les ontologies dans le cadre de l'amélioration de la qualité de la recherche d'informations. Dans la section 3 nous détaillons

notre méthode basée sur l'interprétation sémantique des requêtes utilisateurs par l'usage des ontologies et finalement nous terminerons par une conclusion.

2. Etat de l'art

Une recherche simple qui se base sur des mots clés se heurte à plusieurs limites liées à la variation linguistique, sémantiques et morphologique qui peuvent se présenter dans les composants de la requête utilisateur [Baziz, 02]. Cette variation réduit considérablement l'efficacité des SRI.

L'expansion des requêtes utilisateurs par l'usage des ontologies était l'une des solutions proposées par le Web sémantique pour remédier à ce problème.

Des expériences sur l'expansion des requêtes par des termes reliés sémantiquement ont déjà été effectuées, parmi ces expériences nous citons:

[Gaurino & al, 99] ont utilisé les synsets de WordNet dans le système Ontoseek, pour désambiguïser les termes de requêtes sur les catalogues de produits et les pages jaunes en sélectionnant manuellement les synsets de wordNet [WN] appropriés et leurs catégories.

[Baziz, 02] montre que l'utilisation des ontologies apporte une amélioration notable de la précision dans les résultats de recherche. en s'appuyant sur l'ontologie wordNet, Il à constaté que la relation hyperonymie permet d'améliorer la précision moyenne, alors que la synonymie améliore la précision pour les premiers documents restitués.

[Gonzalo, 98] a proposé une méthode d'indexation des documents s'appuyant sur les concepts d'une base de données sémantique qui améliore la précision lors de la recherche de 25 %.

[Navigili &al 03] ont exploité l'ontologie wordNet dans le processus d'expansion des requêtes et ont montré que l'utilisation des synonymes et des hyperonymes ne contribue pas à une amélioration considérable dans les résultats de recherche. Ils proposent l'utilisation d'autres types d'informations dérivés de l'ontologie. Une requête peut être enrichie par des mots qui figurent dans les définitions des mots de la requête initiale à partir des glossaires de ces derniers. Les auteurs éprouvent que cette méthode apporte une amélioration de 26% dans les résultats de recherche.

Quelque soit la technique adoptée pour l'expansion des requêtes, les auteurs se mettent d'accord sur la nécessité de désambiguïser les sens des termes dans les requêtes initiales par l'utilisation des synsets de WordNet, ceci permet de limiter la taille de la requête étendue, les requêtes longues deviennent vite bruitées et dégradent la précision.

3. Interprétation sémantique des requêtes utilisateurs

3.1 Principe de la méthode :

Commençons par rappeler que nous intéressons aux documents pédagogiques dans le domaine informatique.

Nous partons donc d'une requête à l'état brut qui est formulée par un utilisateur voulant chercher des documents pédagogiques. Cette requête va subir un processus de raffinement pour dissocier le type de document recherché de son contenu. En effet, dans SYFAX (notre plateforme expérimentale), les types des documents sont connus, ce qui facilite leurs énumérations. Une fois le type de ressources recherchées connu, nous pouvons utiliser l'ensemble des mots clés de la requête pour la recherche dans l'entrepôt de SYFAX des documents répondant à ces mots clés. Nous nous basons dans cette phase aussi à une ontologie du domaine qui nous permet d'enrichir le vocabulaire utilisé dans la requête pour étendre le champ de recherche.

Schéma de la stratégie :

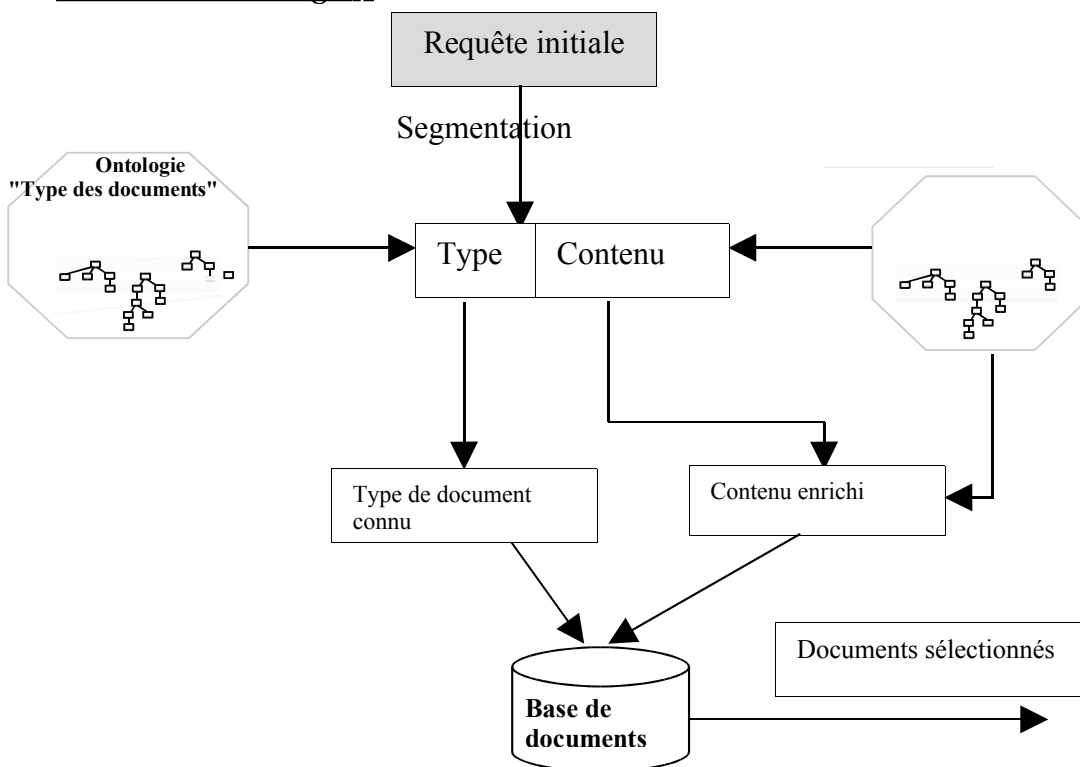


Figure R-1 : Processus de recherche de documents pédagogiques.

3.2 Identification des concepts

a. Identification du type de documents

Comme nous avons déjà expliqué plus haut, nous procédons à une décomposition de la requête initiale de l'utilisateur afin de dissocier le type des documents recherchés du reste des mots clés utilisés. Pour pouvoir identifier le ou les types de documents souhaités par l'utilisateur nous utilisons l'ontologie "type de documents" : une ontologie que nous avons créée manuellement comportant les différents types de documents.

Le processus d'identification commence par former des combinaisons à partir des différents mots constituant la requête initiale. Puis confronter chaque combinaison trouvée à l'ontologie "type de documents" pour tenter d'identifier le type de documents désirés par l'utilisateur.

Exemple :

Examinons le cas de la requête suivante : " Travaux dirigés SQL "

Les Combinaisons possible générées à partir de cette requête sont :

→Travaux

→Dirigés

→SQL

→Travaux dirigés

→Travaux dirigés SQL

→Dirigés SQL

→Travaux dirigés SQL

En confrontons ces combinaisons à l'ontologie "type des documents ", la seule combinaison à retenir dans ce cas est : "Travaux dirigés" qui représente le type de documents souhaité par l'utilisateur.

b. Identification des concepts du domaine

Les utilisateurs cherchent des documents pédagogiques dans le domaine informatique ; donc les mots clés des requêtes sont en quelque sorte des concepts du domaine informatique. Pour détecter ces concepts nous exécutons le même processus adopté pour l'identification du type des documents, mais dans ce cas nous utilisons une ontologie pour le domaine informatique.

L'ontologie que nous avons utilisé est construite automatiquement à partir d'un dictionnaire informatique nommé FOLDOC [FOLDOC] (figure R-2) et ceci en utilisant le système Mecureo [Mecureo], c'est un système créé par Trent Apted [Apted] permettant de générer automatiquement une ontologie dans le domaine informatique à partir du dictionnaire FOLDOC, ce système offre également un module permettant de lancer des requêtes sur l'ontologie.

L'ontologie générée est un graphe pondéré dont les nœuds représentent des concepts du domaine informatique et les liens représentent les relations sémantiques entre les concepts.

```
SQL

<language, database, standard> /S Q L/ An industry-standard
language for creating, updating and, querying {relational
database management systems}.

SQL was developed by {IBM} in the 1970s for use in {System R}.
It is the {de facto standard} as well as being an {ISO} and
{ANSI} {standard}. It is often embedded in general purpose
programming languages.

The first SQL standard, in 1986, provided basic language
constructs for defining and manipulating {tables} of data; a
revision in 1989 added language extensions for {referential
integrity} and generalised {integrity} {constraints}. Another
revision in 1992 provided facilities for {schema} manipulation
and {data administration}, as well as substantial enhancements
for data definition and data manipulation.
```

Figure R-2 : Extrait du dictionnaire FOLDOC représentant la définition du mot "SQL".

4.2 Expansion des requêtes utilisateurs :

Nous avons opté pour une expansion automatique des requêtes utilisateurs. Après avoir détecté les concepts du domaine informatique, Chaque concept est enrichi par des concepts qui sont sémantiquement les plus proches de lui. Pour ce faire on va utiliser le module Foldoccmd du système Mecureo.

Ce module permet de lancer des requêtes sur l'ontologie en recevant comme paramètres de la requête : le concept sujet de la requête, la profondeur souhaité. ce dernier paramètre permet de limiter la taille du graphe résultant de la requête . (la figure R-3 représente un exemple d'une requête lancé sous Foldoccmd pour le concept "SQL" avec une profondeur égale à 5).

le graphe obtenu est formé par le concept en question et les nœuds (les concepts) qui sont les plus proches de lui, ce graphe peut être représenté dans un format RDF [RDF] (la figure R-4 représente le résultat de la requête lancé sous Foldoccmd dans un format RDF) ou dans le format DOT [DOT] .(figure R-5 représente le résultat de la requête lancé sous Foldoccmd "gif" généré à partir du format DOT).

```

C:\mec2>java foldoccmd.QuickDot "ontologie.fdg" SQL 5 >exemple1.DOT
Loading dictionnaire.fdg
10%...
20%...
30%...
40%...
50%...
60%...
70%...
80%...
90%...
100%
Read in 13690 nodes.
Read in 177213 links.
Query size is 5 nodes
Outputting DOT...

```

Figure R -3 exemple d'une requête exemple

d'une requête lancé sous Foldoccmd pour le concept "SQL" avec une profondeur égale à 5

```

<gmp:peer gmp:peerType="child"
rdf:resource="http://foldoc.doc.ic.ac.uk/foldoc/foldoc.cgi?query=SQL2"/>
<gmp:peer gmp:peerType="parent"
rdf:resource="http://foldoc.doc.ic.ac.uk/foldoc/foldoc.cgi?query=Microsoft%
20SQL%20Server"/>
  <gmp:peer gmp:peerType="child"
rdf:resource="http://foldoc.doc.ic.ac.uk/foldoc/foldoc.cgi?query=Red%
20Brick%20Intelligent%20SQL"/>
<gmp:peer gmp:peerType="child"
rdf:resource="http://foldoc.doc.ic.ac.uk/foldoc/foldoc.cgi?query=PostgreSQL
"/>
  <gmp:peer gmp:peerType="sibling"
rdf:resource="http://foldoc.doc.ic.ac.uk/foldoc/foldoc.cgi?query=database%
20query%20language"/>
  <gmp:peer gmp:peerType="synonym"
rdf:resource="http://foldoc.doc.ic.ac.uk/foldoc/foldoc.cgi?query=Structured
%20Q

```

Figure R-4 Résultat de la requête lancé

sous Foldoccmd dans un format RDF

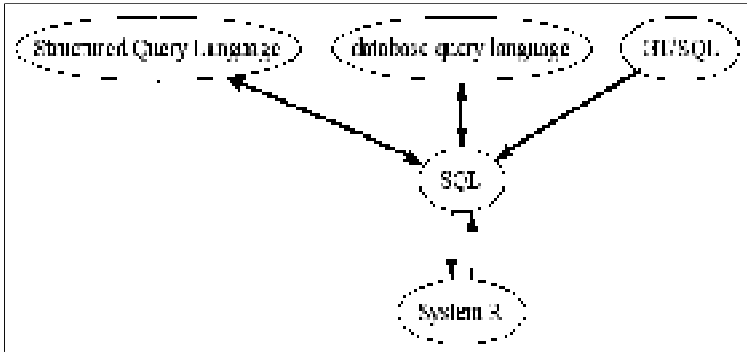


Figure R-5 résultat de la requête lancé sous Foldoccmd dans Format "gif".

5. Conclusion et perspectives

Dans cet article nous avons proposé une nouvelle méthode basée sur l'interprétation sémantique des requêtes utilisateurs dans le cadre d'un système de recherche d'informations. Dans l'objectif d'améliorer la précision du processus de recherche nous avons utilisé deux ontologies : une ontologie "type de documents" et une ontologie du domaine informatique que nous avons exploitée pour enrichir les requêtes utilisateurs.

Cependant, Nous tenons à signaler que :

L'ontologie du domaine informatique généré à partir du système Mecureo à partir de FOLDOC n'est pas complète. Nous comptons l'enrichir par de nouveaux concepts comme :

"Recherche d'informations ", "Web sémantique", etc...

Nous comptons aussi utiliser cette ontologie dans le processus d'indexation des documents pédagogiques ceci permettra une amélioration considérable des résultats de recherche.

Références :

[Apted] <http://www.ug.cs.usyd.edu/~taped/>.

[Aussenac, 02] N. Aussenac, Support de cours conçu par N. Aussenac-Gilles, J. Charlet, P. Laublet et B. Bachimont. Cours sur les Ontologies, les Terminologies et les Bases de Connaissances Terminologiques : <http://www.irit.fr/GRACQ> , (2002).

[Baziz, 02] M. Baziz, « Application des Ontologies pour l'Expansion de Requêtes dans un Système de Recherche d'Informations », Rapport de DEA 2IL Irit, (juin 2002).

[DOT] an open source graph visualisation tool, at <http://www.research.att.com/sw/tools/graphviz/>.

[FOLDOC] the free On-Line Dictionary Of Computing[© 1993 by Denis Howe, updated regularly], at <http://foldoc.doc.ic.ac.uk/foldoc/contents.html>.

[Mecureo] <http://www.it.usyd.edu.au/~taped/projects.html#Mecureo>

[Guarino & al, 99] Nicola Guarino, Claudio Masolo, and Guido Vetere. "OntoSeek : contentbased access to the web". *IEEE Intelligent Systems*, (1999).

[Gonzalo, 98] Julio Gonzalo, Felisa Verdejo, Irina Chugur, and Juan Cigarran. Indexing with wordnet synsets can improve text retrieval. In Proceedings of the COLING/ACL '98 Workshop on Usage of WordNet for NLP, pages 38-44, Montreal, Canada, (1998).

[Navigli et al 2003]. An Analysis of Ontology-based QueryExpansion Strategies. Roberto Navigli and PaolaVelardi. Workshop on Adaptive Text Extraction and Mining (ATEM2003), in the 14th European Conference on Machine Learning(ECML 2003).

[RDF] Resource Description Framework, at <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

[SFX1] H. SMEI, M. MAKPANGOU, A. BEN HAMADOU, SYFAX : Un système de cache Web Sémantique pour des communautés Distribuées, Conférence Internationale – MediaNet 2002 – Juin 2002 – Sousse – Tunisie

[SFX2] H. SMEI, M. MAKPANGOU, A. BEN HAMADOU, Vers un SYstème de Fouille Automatique et d'eXtraction d'information : SYFAX, Cas de l'enseignement. GEI 2002 – Mars 2002 – Hammamet – Tunisie.

[WN] <http://www.cogsci.princeton.edu/~wn/>.

