A QUERY EXPANSION METHODOLOGY IN A COOPERATION OF INFORMATION SYSTEMS BASED ON ONTOLOGIES

Guillermo Valente Gómez Carpio, Lylia Abrouk and Nadine Cullot
LE2I, UMR CNRS 5158, University of Burgundy, Dijon, France

Keywords: Query expansion, Ontology, Cooperation of Information Systems.

Abstract: The great development of Internet technologies and the emergence of the semantic web have led to the specification of architectures and tools to describe and to allow the "relevant" sharing of heterogeneous information sources. Shared data can be annotated and mapped to an agreed representation of their semantic using ontologies. Ontology is a representation of a domain of interest which is agreed by a community of people. The aim of the paper is to propose a user's query expansion methodology in a heterogeneous cooperation of information systems. We propose a complete architecture called OWSCIS (Ontology and Web Services Cooperation of Information Sources) of cooperation of information systems based on the use of reference ontology and several local ontologies. This paper highlights the query expansion methodology in this architecture. The objective is to help and guide the user during the query process by analysing his query and using the usual behaviours of the users to predict his need.

1 INTRODUCTION AND MOTIVATION

The emergence of the semantic web as a mean to share "understandable" data from heterogeneous sources is still a great challenge. The main idea is to be able to annotate and map information sources data to an ontology to express their semantic and to transparently query the cooperation.

- *Information System Cooperation*. Lots of works have been developed from classical databases information systems cooperation to more dynamic cooperation of information sources. We can note the evolution of the information systems cooperation architectures from federated to mediator-wrapper architectures and some more semantic-based architectures using ontologies to express the understanding of the shared knowledge.
- *Ontologies.* They can be viewed as an agreed representation of a domain of interest for a community of people, usable to share data. The information sources can be annotated and mapped to one or several ontologies to describe their semantic allowing the transparent querying of the cooperation.
- *Cooperation Querying.* The aim of cooperation is to give users the opportunity to transparently query the cooperation. This querying can be done using the agreed representation of the knowledge i.e. the ontology.

We can mention some examples of relevant systems which aim at defining such cooperation environment such as MOMIS (Beneventano and Bergamaschi, 2004), MECOTA (Stuckenschmidt and Wache, 2000), Infosleuth (Nodine et al., 2000) or Kraft (Preece et al., 2000).

The transparent querying of a cooperation raises a lot of complex problems such as: (i) the choice of the language to pose the query (natural language, list of terms, formal language, etc.), (ii) the query decomposition to access to the data of the different sources of the cooperation, (iii) the results re-composition to answer the initial query and also (iv) the need to help and guide the users to build their queries.

To motivate our work on query expansion methodology, we first present its context of use with a brief presentation of our architecture of cooperation called OWSCIS (Ontology and Web Service based Cooperation of Information Sources).

Carpio G., Abrouk L. and Cullot N.
A QUERY EXPANSION METHODOLOGY IN A COOPERATION OF INFORMATION SYSTEMS BASED ON ONTOLOGIES.
DOI: 10.5220/0001825302560262
In *Proceedings of the Fifth International Conference on Web Information Systems and Technologies (WEBIST 2009)*, page
ISBN: 978-989-8111-81-4

1.1 OWSCIS Architecture Overview

The OWSCIS architecture (Ontology and Web Service based Cooperation of Information Sources), is based on ontologies and web services technologies to allow the cooperation of distributed and heterogonous information sources. In this architecture, information sources are encapsulated in modules called "Data Providers". An information source may be a relational database or an XML document, and a data provider may contain one or more information sources. A data provider wraps its information source(s) to a local ontology which expresses the semantic of information sources. Local ontologies are also mapped to a reference ontology that semantically describes a given domain of interest and covers all participant local ontologies. Most of the architecture components are encapsulated in web services, each of them performs a specific task including: inter-ontology mapping, query processing and query-results visualization.

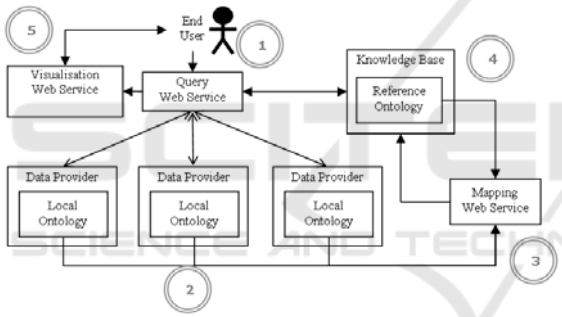

Figure 1: OWSCIS Architecture.

Figure 1 gives an overview of this architecture. The query web service allows end users to submit their queries in terms of the reference ontology. The query is expressed in *sparql* language; it is analyzed and decomposed into a set of sub-queries over the relevant data providers. The results returned from the different data providers are recomposed and sent to the visualization web service. The sparql query is also analyzed to be enriched during the querying process.

The following section gives an overview of the query expansion method which is applied to help and guide users during the querying process.

It is important to note that users directly query the cooperation through its reference ontology and that the answers returned to the users are enriched with related terms of the ontology. The query expansion methodology we propose is a mean to choose pertinently these terms regarding the usual behaviors of the users.

1.2 Expansion Method Overview

The aim of the query expansion method is to enrich the query of a user by providing him the results of his initial query but also a list of terms associated to the terms of his initial query. These terms may be used by the user to refine his query or to build a new query. In particular, in our solution, the related terms are chosen by analyzing the usual behaviour of the users.

Figure 2 gives an overview of query expansion process applied in the system OWSCIS.

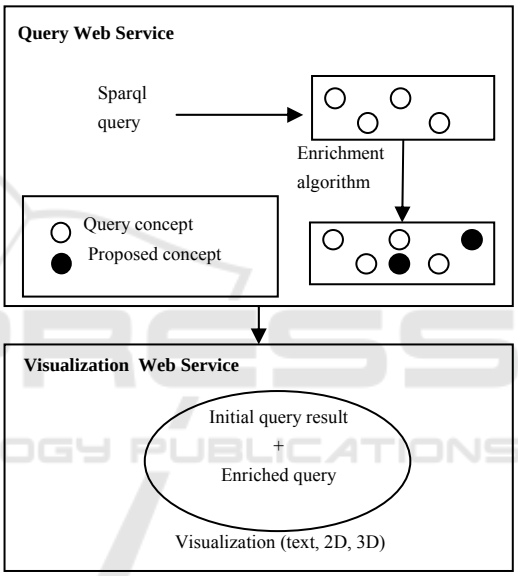

Figure 2: OWSCIS Query Expansion Process.

This query expansion process consists in four steps: 1) the user submits his query (in SPARQL) on the reference ontology, 2) the query elements (concepts and properties) are extracted to build a list of terms to be enriched, 3) the query expansion algorithm is applied to enrich the initial list of the concepts extracted from the query with a list of proposed concepts also belonging to the reference ontology and 4) the user can visualize the results of his initial query but also the enriched query.

The paper is organized as follows. Section 2 proposes a discussion on some existing expansion methods based on different approaches. Section 3 is dedicated to the details of the proposed query expansion methodology. Section 4 illustrates the methodology and section 5 concludes this paper.

2 QUERY EXPANSION METHODS: RELATED WORKS

Query expansion in information retrieval systems attempts to make answers more clear and precise. This allows users to modify their queries and improves retrieval performance (Salton and McGill, 1986).

Several works have tackled this problem. The approaches described below are based on (i) documents retrieval, (ii) statistical methods and (iii) ontology-based methods by calculating a semantic similarity, etc.

2.1 Query Expansion Methods using Documents

Raymond et al., (2002) present a query reformulation system based on an unsupervised classification method. The information retrieval system called SIAC is used to obtain a set of initial documents from a query and then each document is associated to a set of terms. The system classifies the sentences of the retrieved document according to some words extracted from the document.

Other query expansion methods are based on the use of the profile of the user.

Bottraud and G. Bisson, (2004) propose an approach which is based on the building of the user's profile in order to enrich his queries. The system identifies a context using the user's profile and his query. The query expansion is done in two phases: (i) vocabulary learning and (ii) queries learning. The profile of the user is updated according to his search.

Another approach (Peninou et al., 2006) is also based on the use of user's profile and consists in comparing the similarity between the query and the user's profile elements.

2.2 Statistical Query Expansion Methods

Some statistical methods are generally based on the notion of co-occurrence of terms. Terms are grouped into classes and the terms of a same class are used to enrich the queries. It consists in creating relations between the terms of a same document.

Cui et al., (2002) assume that the terms occurring in a query are correlated to the terms of the document on which the user has clicked. The expansion method is based on a probabilistic analysis. It extracts correlations among the terms of the query and the documents by analyzing the query logs. The best terms from documents *(high quality)* are used to expand the query.

Instead of calculating similarity with each term of the query, (Qiu and Frei, 1993) take into account the whole query and consider terms similar to the "query concept".

2.3 Semantic Query Expansion Methods

According to Voorhees, (1994), statistical methods are less successful than methods using semantic and linguistic relations. He proposes an enrichment query method based on the use of the semantic relations between the terms (synsets) as defined in Wordnet. In order to enrich a query, several methods are also based on the use of external resources like ontologies. The relations between terms are used to expand the terms of the query.

Tomassen et al., (2006) use an ontology to expand the queries and improve the quality of the information retrieval. They use text mining techniques to adapt the ontology concepts to the domain terminology (terms used in documents but not in query).

Schweighofer and Geist, (2007) combine two techniques: the use of ontologies and the relevance feedback. The terms of the query are looked up in a knowledge base (lexical ontology) and weighted.

2.4 Discussion

Most of the approaches described in the previous section, propose methods which enable the expansion of user queries either using the users profile or terms semantically "close" to the terms of queries. In the latter case, the way to obtain the "similar" terms can differ. They can be found using some statistical methods by analyzing the co-occurrence of terms in some documents and in the query or using external resources such as thesaurus or more complex ontologies. Few approaches use the relevance feedback which allows the consideration of the actions "really" done by the users.

We propose a query expansion method which is based on the analysis of the usual behaviors of the users. The initial and the enriched queries are expressed using terms of a reference ontology.

The proposed method can be viewed as a variant of the PageRank algorithm used to find and rank web pages. However, our algorithm is applied in a

different context. The notion of link between pages used in the page rank algorithm is substituted by the notion of usual behavior of the users in their queries and the value of the page rank for a page is replaced by the notion of ConceptRank value which expresses the popularity of a concept and by the notion of importance of concept relatively to another one.

We can note that some works such as those proposed by (Richardson and Domingos, 2002) focus on the improvement of the PageRank algorithm with a probabilistic model to find the relevant pages. Our methodology differs from these approaches because it is based on the analysis of the concepts of the queries and the behavior of the users.

3 THE QUERY EXPANSION METHOD

As we discuss below, our method is based on the page rank algorithm but the value of the page rank for a page is replaced by the notion of *ConceptRank value which expresses the popularity and the importance of a concept relatively to another one*.

The basic version of PageRank algorithm is based on the idea that the more a document is important, the more other documents are linked to it.

We apply this heuristic to concepts with the idea that the more a concept is important, the more this concept is searched after another concept in the users' queries.

The PageRank algorithm is based on the analysis of the hypertext graph of the links between the pages. In our method, we build a concept graph which is based on the terms successively used by the users in their queries.

3.1 Concept Graph Definition

A concept graph is built on the basis on the users' queries. Intuitively, the nodes of the graph are the concepts and an arc from a node N1 to a node N2 expresses that the concept of the node N2 has been searched after the concept of the node N1 by a user. This graph is directed and weighted.

First, we introduce the definitions of a query sequence and sub-sequence, then we define the notions of concept graph for a sub-sequence and a sequence of queries.

Definition. Let C be a set of concepts of a domain, Q a query which is a subset of concepts of C, a query sequence is a n-uplet $S(Q_1,Q_2,...,Q_n)$ where Q_i is a query.

A query sub-sequence of a sequence $S(Q_1,Q_2,...,Q_n)$ is a 2-uplet $S_i(Q_i,Q_{i+1})$ where $i=1\ldots n-1$.

The cardinality $|Q|$ of a query Q, is the number of concepts belonging to this query.

Figure 3 illustrates a query sequence $S(Q_1,Q_2,Q_3)$. Each sub-sequence Si of a sequence S is represented by a concept graph G_i.

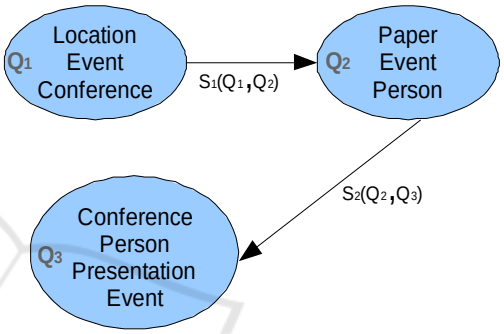

Figure 3: A Query Sequence $S(Q_1,Q_2,Q_3)$.

Definition. A concept graph $G_i(O_i,E_i)$ is a directed weighted graph, for a sub-sequence $S_i(Q_i,Q_{i+1})$ of a sequence $S(Q_1,...,Qn)$ where:

O_i is the set of the concepts belonging to $Q_i \cup Q_{i+1}$,

V_i is the set of edges belonging to

$(Q_i \cap Q_{i+1}) \times (Q_{i+1} - Q_i)$.

The weight $w_i(v_i)$ associated to an edge $v_i \in V_i$ is defined by:

$$w_i(v_i) = \frac{1}{|Q_i \cap Q_{i+1}|} \quad (1)$$

The value $w_i(v_i)$ represents the weight associated to an arc. It is based on the number of concepts that a user has kept between two successive queries of a sub-sequence $S_i(Q_i, Q_{i+1})$.

Definition. A graph G *for a sequence* $S(Q_1,...,Qn)$ is the union of the G_i graphs of the sub-sequences $S_i(Q_i,Q_{i+1})$.

$$G = \bigcup_{i=1}^{n-1} G_i \quad (2)$$

Figure 4 illustrates the concept graphs G_1 and G_2 for the sub-sequences $S_1(Q_1,Q_2)$ and $S_2(Q_2,Q_3)$ of the sequence $S(Q_1, Q_2,Q_3)$ shown in figure 3.

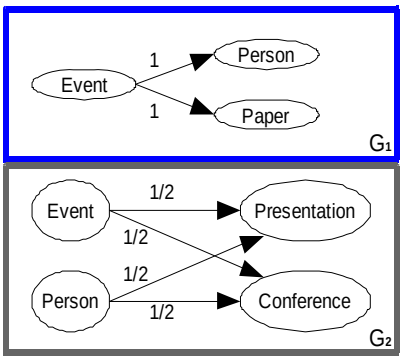

Figure 4: Concept graphs of the sub-sequences.

The concept graph of a query sequence is translated into a matrix. Intuitively, this matrix contains for each couple of concepts (C1, C2), the sum of the weights of the arcs existing between the node C1 and C2 in the concept graph.
We first define below the notion of matrix for a sub-sequence and then the notion of matrix for a whole sequence.

Definition. A graph G_i for a sub-sequence $S_i(Q_i,Q_{i+1})$, can be represented by a matrix MC_i, such as:
$MC_i : C \rightarrow C$, where C is the set of concepts of the domain.

$$MC_i(c_i,c_j) = \begin{cases} w_i(c_i,c_j) \text{ if there is an arc } v_i: c_i \rightarrow c_j \\ 0 \text{ else.} \end{cases} \quad (3)$$

Definition. *The matrix MC* of a graph G of a sequence $S(Q_1,...,Qn)$ is the sum of the matrix MC_i of the sub-sequences $S_i(Q_i,Q_{i+1})$.

$$MC = \sum_{i=1}^{n-1} MC_i \quad (4)$$

Definition. We note $N(c)$ the number of arcs of G, going from c to another concept.

$$N(c) = \sum_{j=0}^{n} v_j \quad \begin{cases} v_j=1, \text{ if } MC(c_, c_j)\neq 0 \\ v_j=0, \text{ else} \end{cases} \quad (5)$$

For each query sequence, a concept graph is computed and the matrix is incremented.

3.2 ConceptRank Measure

Based on the matrix defined in the previous section, we can now introduce a new measure of popularity called the ConceptRank measure.

As in the PageRank algorithm, the computation of ConceptRank measure needs a number of iterations to fix the value. The ConceptRank measure expresses the popularity of a concept according to the usual behavior of the users during their querying process. The ConceptRank measure is defined as follows.

Definition. Let c_i be a concept of C, $B(c_i)$ the set of concepts, such that $MC(c_i, c_i)\neq 0$, d a normalization factor:
the ConceptRank measure of a concept c_i ,$CR(c_i)$ is defined by:

$$CR(c_i) = (1-d) + d \sum_{c_j \in B(c_i)} \left(\frac{CR(c_j)}{N(cj)} \right) \quad (6)$$

We complete this notion of popularity of a concept with the notion of the importance of a concept relatively to another concept or to a query.

3.3 Concept Importance Measure

The notion of importance of a concept relatively to another concept and more precisely the importance of a concept relatively to a whole query is the measure that we use to choose the concepts to add to enrich the user's query. These measures are based on the popularity of the concepts as defined previously with the ConceptRank measure.

Definition. Let MC the matrix representing some graph G and Q a query.
$CI(c_i,c_j)$ is the Concept Importance of the concept c_i relatively to the concept c_j which is defined by:

$$CI(c_i,c_j) = MC(c_i,c_j) * CR(c_i) \quad (7)$$

$I(c_i,Q_i)$ is the Importance of a concept c_i relatively to a query Q_i which is defined by:

$$I(c_i, Q_i) = \sum_{c_j \in Q_i} CI(c_i, c_j) \quad (8)$$

3.4 The Minimum Importance Factor

To choose the concepts to add to expand a query considering their importance relatively to this query, we define a threshold value called the minimum factor importance. A query Q_i is expanded with a concept c_i if its concept importance $I(c_i,Q_i)$ is greater than the minimum importance factor.

Definition. Let I_{Qi} be the set of concepts importance values for a query Q_i, $I_{Qi} = \{I(c_1,Q_i),\ldots,I(c_n,Q_i)$.
The minimum importance factor *Min (FI/Q_i)*for a query Q_i is defined by:

$$Min(FI/Q_i) = \left(\frac{Max(I_{Q_i}) + Min(I_{Q_i})}{2} \right) \quad (9)$$

4 EXPERIMENTATION

The presented approach has been implemented. We introduce in this section some preliminary results. To simplify the description of the experimentation, we directly consider the concepts extracted from the SPARQL queries done on the reference ontology of the OWSCIS architecture. The used ontology is a sample "Conference" ontology freely available on the Web.

The experimentation is composed of two phases:

- *Learning phase:* this initial phase consists in storing in the knowledge base a set of query sequences in order to be able to compute the importance of the concepts relatively to a new query. The enrichment process is based on this phase. Then the knowledge base is updated for each new query sequence.
- *Query Expansion:* the user submits a new query and the system returns the results of this initial query and also a list of new concepts related to the initial query. The user can visualize all the concepts in the reference ontology.

The experimentation has been done on 55 users' queries for the learning phase. Then the matrix of concepts (MC) has been calculated. This matrix is incremented for each new query sequence done by a user. We also calculated the ConceptRank values of the concepts and all the necessary values of importance of a concept relatively to a new query to be enriched.

Table 1 gives a short extract of the results with the description of the query to be expanded (first column), the value of the minimum importance factor for this query (second column) and the expansion of the query (third column). The added concepts are in bold. The new concepts occurring in the enriched query are those for which the importance value relatively to the query is greater than the minimum importance factor.

Table 1: Query expansion.

Query	*Min(FI)*	Query Expansion
Conference, Event	1.1205	Conference, Event, **Location, Presentation**
Commitee, Organization	1.304	Commitee, Organization, **Person, SocialEvent, Conference**
Location, SocialEvent	1.57	Location, SocialEvent, **Presentation, Organization**
KeynoteTalk, Presenter, Presentation	1.5515	KeynoteTalk, Presenter, Presentation, **Event, Paper**
Person, Group, Event	2.4965	Person,Group, Event, **Committee, Organization**
Presentation, Conference, Paper	1.685	Presentation, Conference, Paper, **Event**
Organization, Workshop, Presentation	1.467	Organization, Workshop, Presentation, **Paper, Presenter**
Commitee, Location	1.417	Commitee, Location, **Conference, Presentation, Organization**

5 CONCLUSIONS

In this paper, we have presented a new query expansion method based on the use of ontology and a new measure of popularity and importance taking into account the usual behavior of the users. The query expansion process is included in the query and visualization processes of the system OWSCIS (Ontology and Web Service based Cooperation of Information Sources). Some preliminary results are given in section 5. An experimentation on a huge ontology (1500 concepts) is currently performed to confirm the method. The quality and the pertinence

of the results are greatly linked to the "intelligent" behavior of the users. A more detailed analysis of this knowledge to construct the graph is also currently studied to optimize this work.

REFERENCES

Bottraud, J-C., Bisson, G., Bruandet, M-F., 2004. Expansion de requêtes par apprentissage automatique dans un assistant pour la recherche d'information. In *CORIA'04*, *Conférence en Recherche Information et Applications*., pages 89-108.

Beneventano D., Bergamaschi, S., 2004. Methodology for integrating heterogeneous data sources. In *IFIP'04, World Computer Congress*., pages 22-27.

Cui, H., Rong Wen, J., Yun Nie, J., Ying Ma, W., 2002. Probabilistic query expansion using query logs. In WWW'02, *11th international World Wide Web conference.,* pages 325-332. ACM Press.

Nodine, M., Fowler, B., Ksiezyk, T., Perry, B., Taylor, M., Unruh, A., 2000. Active Information Gathering in InfoSleuth, In *International Journal of Cooperative Information Systems*., pages 3-28.

Peninou, C.Z.A., Canut, M., Sedes, F., 2006. An adaptation approach: query enrichment by user profile. In *SITIS'06, The international conference on Signal-Image Technology & Internet-Based Systems.,* pages 24-35.

Preece, A., Hui, K., Gray, A., Marti, P., Bench-Capon, T., Jones, D., Cui, Z., 2000. The KRAFT architecture for knowledge fusion and transformation. In *Knowledge-Based Systems*, Volume 13, Issues 2-3, pages 113-120.

Quiu, Y., Frei, H.-P., 1993. Concept based query expansion. In *SIGIR'93,16th Annual International Conference on Research and Development in Information Retrieval.,* pages 160-169, New York, NY, USA. ACM Press.

Raymond, C., Bellot, P., El-Bèze, M., 2002. Enrichissement de requêtes pour la recherche documentaire selon une classification non-supervisée. In *RFIA'02, 13ème Congrès Francophone AFRIF-AFIA de Reconnaissance des Formes et d'Intelligence Artificielle.*, pages 625-632.

Richardson, M., Domingos, P., 2002. The intelligent surfer : Probabilistic combination of link and content information in PageRank. In *Advances in Neural Information Processing Systems Vol. 14.,* pages 1441-1448, Cambridge, MA: MIT Press.

Salton, G., McGuill, M., 1986. *Introduction to Modern Information Retrieval.* McGraw-Hill, Inc., New York, NY, USA.

Schweighofer, E., Geist, A., 2007. Legal query expansion using ontologies and relevance feedback. In *LOAIT'07, Legal Ontologies and Artificial Intelligence Techniques.,* pages 149-160.

Stuckenschmidt, H.,Wache, H., 2000. Context modeling and transformation for semantic interoperability. In *KRDB'00, Knowledge Representation Meets Databases.*

Tomassen, S., Gulla, J., Stransunskas, D., 2006. Document space adapted ontology: Application in query enrichment. In *NLDB'06, 11th International Conference on Applications of Natural Language to Information Systems*., pages 46-57.

Voorhees, .E. M., 1994. Query expansion using lexical-semantic relations. In *SIGIR '94*, *17th annual international ACM-SIGIR conference on Research and Development in Information Retrieval,* pages 61-69, New York, NY, USA. ACM Press.