Query Expansion and Interpretation to Go Beyond Semantic P2P Interoperability

Anthony Ventresque, Sylvie Cazalens, Philippe Lamarre, Patrick Valduriez

Query expansion and interpretation to go beyond semantic P2P interoperability

Anthony Ventresque[1], Sylvie Cazalens[1], and Philippe Lamarre[1] and Patrick Valduriez[2]

[1]LINA, University of Nantes
FirstName.LastName@univ-nantes.fr
[2]INRIA and LINA, University of Nantes
Patrick.Valduriez@inria.fr

Abstract. In P2P data management systems, semantic interoperability between any two peers that do not share the same ontology relyes on ontology matching. The established correspondences, i.e. the "shared" parts of the ontologies are indeed essential to exchange information. But to what extent the "unshared" part can contribute to information exchange. In this paper, we address this question. We focus on a P2P document management system, where documents and queries are represented by semantic vectors. We propose a specific query expansion step at the query initiator's side and a query interpretation step at the document provider's. Through these steps, unshared concepts contribute to evaluate the relevance of documents wrt a given query. The experiments show that the proposed method enables to correctly evaluate the relevance of a document even if concepts of a query are not shared. In some cases, we are able to find up to 90% of the documents that would be selected when all the concepts are shared.

1 Introduction

In peer-to-peer (P2P) data systems, semantic interoperability means that any two peers are able to exchange information of which meaning is correctly interpreted by both of them. Several solution in P2P data management use local mappings to ensure the systems global interoperability [5, 8]. Most of the solutions focus on what (i.e. the concepts and relations) the peers share, which is important. However, no matter how the shared part is obtained (through consensus or mapping), there might be concepts (and relations) that are not consensual, and thus not shared but still useful for information exchange. Thus the question is to know whether the unshared parts are useful for information exchange.

In this paper, we restrict this question to the case of a P2P document management system, with unstructured or semi-structured documents. More precisely, we focus on semantic interoperability and information exchange between two peers, a query initiator p_1 and a document provider p_2, which use different ontologies but share some common concepts. Each of them represents its queries

and documents, according to its own ontology. The, the problem we address is to *find documents which are relevant to a given query although the documents and the query may be both represented with concepts that are not all shared.*

We represent documents and queries by *semantic vectors* [11] which are a common way to represent unstructured documents. The principle is simple: each concept of the ontology is weighted according to its representiveness of the document. The same is done for the query. The resulting vector represents the document (respectively, the query) in the n-dimensional space formed by the n concepts of the ontology. Then the relevance of a document with respect to a query corresponds to the proximity of the vectors in the space.

In order to improve information exchange beyond the "shared part" of the ontology, we promote both query expansion (at the query initiator's side) and query interpretation (at the document provider's side). Query expansion may contribute to weight linked shared concepts, thus improving the document provider's understanding of the query. Similarly, by interpreting an expanded query with respect to its own ontology (i.e. by weighting additional concepts of its own ontology), the document provider may find additional related documents that would not be found by only using the matching concepts in the query and the documents. To our knowledge, the best of the problem of improving information exchange by using the unshared concepts of different ontology has not been addressed before. Our proposal is a first, encouraging solution.

This paper is organised as follows. Section 2 gives preliminary definitions. In Section 3, we present query expansion in the case of a shared ontology. Its main property is to keep separate the results of the propagation from each central concept of the query. Section 4 considers the case where two peers use different ontologies, and describes query interpretation. Section 5 gives preliminary experiments. Finally, we conclude in Section 6.

2 Preliminary Definitions

We use a semantic vector space, i.e. a multi-dimensional linear space with the concepts of an ontology as dimensions. The content of each document (respectively query) is abstracted to a semantic vector by characterizing it according to each concept. The more a given document is related to a given concept, the higher is the value of the concept in the semantic vector of the document.

We simply define an ontology as a set of concepts together with a set of relations between those concepts [4]. In our experiments we consider an ontology with only the is-a relation (specialization link). This does not restrict the generality of our relevance calculus.

Definition 1 (Semantic Vector). *A* semantic vector $\overrightarrow{v_\Omega}$*, or* $\overrightarrow{v}$ *when there is no ambiguity, is an application defined on the set of concepts* $\mathcal{C}_\Omega$ *of an ontology* Ω*:*

$\forall c \in \mathcal{C}_\Omega, \overrightarrow{v_\Omega} : c \rightarrow [0, 1]$

Expansion is based on weight propagation, which consists in weighting initially unweighted concepts which seem linked to weighted concepts. We propose to propagate weight from a given concept c_i , according to the similarity[7] of the other concepts with c_i . Thus, we introduce a similarity function sim_{c_i} which denotes the similarity to a *central concept* c_i .

Definition 2 (Similarity function). *Let c be a concept of $\mathcal{C}_\Omega$. Function sim_c: $\mathcal{C}_\Omega \rightarrow [0,1]$, is a similarity function iff $sim_c(c) = 1$ and $0 \leq sim_c(c_j) < 1$ for all $c_j \neq c$ in $\mathcal{C}_\Omega$.*

The concepts of $\mathcal{C}_\Omega$ can then be ordered according to their similarity value. A propagation function $\mathcal{P}f_c$ is a decreasing function which assigns a weight to each similarity value, c being assigned the highest weight. Figure 1 is an example of a propagation function, inspired by membership functions used in fuzzy logic.

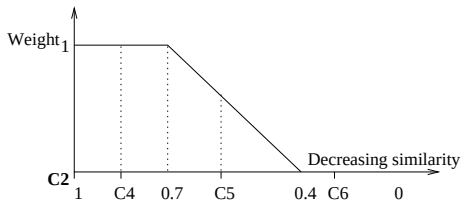

Fig. 1. Example of a $f_{0.7,0.4,1}$ function with central concept c_2

3 Query Expansion and Image based Relevance

For the sake of simplicity, we assume in this section that the query initiator and the document provider use the same ontology but they can still differ on the similarity measures and the propagation functions.

Most propagation methods propagate the weight of each concept in *the same vector*. We call this kind of method "rough" propagation. Although the results are not bad, this method has some drawbacks, in particular, a possible unbalance of the relative importance of the initial concepts [6]. This is why we choose to keep separate the results of the propagation from different concepts in *semantic enriched dimensions* (SEDs).

First, let us denote by $\mathcal{C}_{\overrightarrow{q}}$ the set of the *central concepts* of query $\overrightarrow{q}$, i.e. those weighted concepts which best represent the query. Each central concept of $\mathcal{C}_{\overrightarrow{q}}$ is *semantically enriched* by propagation, in a separate vector (see figure 2).

Definition 3 (Query expansion). *Let $\overrightarrow{q}$ be a query vector; let c be a concept in $\mathcal{C}_{\overrightarrow{q}}$ and let $\mathcal{P}f_c$ be a propagation function.*
The semantically enriched dimension of c, *noted* $\overrightarrow{sed}_c$, *is the semantic vector*

defined by: $\forall c' \in \mathcal{C}_\Omega, \overrightarrow{sed}_c[c'] = \mathcal{P}f_c(c')$
The expansion *of* $\overrightarrow{q}$*, noted* $\mathcal{E}_{\overrightarrow{q}}$ *is defined as:* $\mathcal{E}_{\overrightarrow{q}} = \{\overrightarrow{sed}_c : c \in \mathcal{C}_{\overrightarrow{q}}\}$

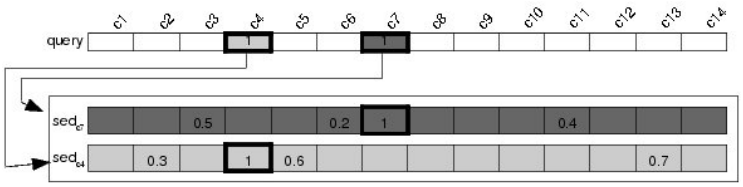

Fig. 2. A query expansion composed of 2 semantically enriched dimensions.

The relevance of a given document is computed using the cosine of its *image* wrt the query and the query itself. This image is obtained using the expansion of the query (i.e. the set of SEDs). Given a SED $\overrightarrow{sed}_c$, we consider the product of the respective values of each concept in $\overrightarrow{sed}_c$ and $\overrightarrow{d}$. The image of $\overrightarrow{d}$ keeps track of the best value assigned to one of the linked concepts if it is better than $\overrightarrow{d}[c]$, which is the initial value of c. All the central concepts of the initial query vector are then weighted in the image of the document as far as the document is related to them.

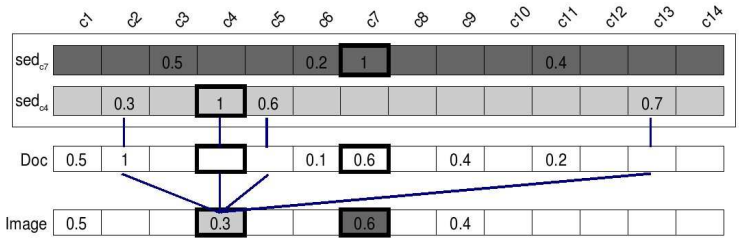

Fig. 3. Obtaining the image of a document.

4 Relevance in the Context of Unshared Concepts

We now assume that the query initiator, p_1, and the document provider, p_2, use different ontologies, respectively noted Ω_1 and Ω_2. Each peer also has its own similarity and propagation functions. We also assume that the peers *share* some common concepts: each of them regularly (although may be not often) computes an ontology matching algorithm which provides a non-empty set of

correspondences (equivalences) between those concepts [3]. For the sake of simplicity of notations, when there is an equivalence, we make no difference between the name of the given concept at p_1's, its name at p_2's, and the identifier of the correspondence, which all refer to the same concept.

4.1 Overview of the Relevance Calculus

The query initiator and the document provider do not use the same vector spaces. An additional step is needed in order to be able to evaluate relevance in a same and single space. We call it *interpretation* of the query. Thus, the different steps involved in the relevance calculus of some document $\overrightarrow{d}$ of p_2 wrt a given query $\overrightarrow{q}$ initiated by p_1 are the following.

Query expansion. It remains unchanged. Peer p_1 computes an *expansion* of its query, which results in a set of SEDs. Each SED is expressed on the set $\mathcal{C}_{\Omega_1}$, no matter the ontology used by p_2. Then, the expanded query is sent to p_2, together with the initial query.

Query interpretation. Query interpretation by p_2 provides a set of interpreted SEDs on the set $\mathcal{C}_{\Omega_2}$ and an interpreted query. Each SED of the expanded query is interpreted separately. Interpretation is composed of two problems:

- The first problem is to find a concept that corresponds to the central concept of the SED. This is difficult when the central concept is not shared. It might even lead p_2 to introduce "new" concepts. Because of space limitations, we do not detail this part. In the following, we assume that the corresponding concept belongs to $\mathcal{C}_{\Omega_2}$, even if the initial concept is not shared, and that it keeps the weight of this latter.
- The second problem is to attribute weights to unshared concepts of $\mathcal{C}_{\Omega_2}$ which are linked to the SED. This is detailed below.

Image of the document and cosine calculus. They remain unchanged. Provider p_2 computes the image of its document with respect to the interpreted SEDs and then, its cosine based relevance with respect to the interpreted query, no matter the ontology used by p_1. This is possible because the previous interpretation step makes both the image of the document and the interpreted query belong to space $\mathcal{C}_{\Omega_2}$.

4.2 Interpretation of a SED

In this section, we describe the interpretation process for a given SED (expressed on $\mathcal{C}_{\Omega_1}$), of which central concept is noted c. The concept corresponding to c in $\mathcal{C}_{\Omega_2}$ is noted $i_{\mathcal{E}\overrightarrow{q}}(c)$ and is assigned the weight of c. Peer $p2$ ranks its own concepts in function of $sim_{i_{\mathcal{E}\overrightarrow{q}}(c)}$. Among these concepts, some are shared and their initial SED has a given weight, which we preserve in the interpretation. The problem is how to weight the unshared concepts, given that some of them might be more similar to $i_{\mathcal{E}\overrightarrow{q}}(c)$ than shared concepts. Figure 4 illustrates our

general solution. Let us call f_i a piecewise affine function which defines a weight for each similarity value in $[0, 1]$. To guide the definition of f_i, we use the weights given by $\overrightarrow{sed}_c$ to the shared concepts ($c1$, $c2$, $c3$ in figure 4). However, there might be several shared concepts that have the same similarity value with respect to $i_{\mathcal{E}_{\overrightarrow{q}}}(c)$, but have a different weight according to $\overrightarrow{sed}_c$. We only require function f_i to assign one of (or a function of) these values to the given similarity value. For instance, it can be the minimum value. Given a function f_i that satisfyes this condition, the unshared concepts ($c4$, $c5$, $c6$ in the figure) are assigned the weight they obtain by function f_i. This is illustrated for c_5 by a dotted arrow.

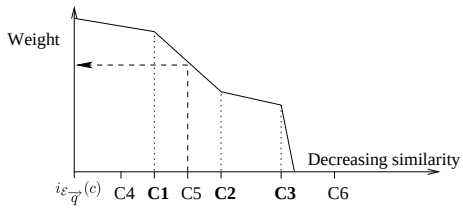

Fig. 4. SED interpretation: assigning weights to unshared non-central concepts

5 Preliminary Experiments

The ontology we use is lightweight, i.e. an ontology composed of a taxonomy of concepts and a taxonomy of relations : *WordNet*. We use the *Cranfield corpus*, a testing corpus consisting of 1400 documents and 225 queries in natural language, all related to aeronautical engineering[1]. For each query, each document is scored by humans as relevant or not relevant (boolean relevance). Although it is similar in size to other classical testing corpora like CACM, CISI, Medline, etc. it is small compared to recent TREC corpus. As we are not experts in textual IR nor in natural language processing, we do not focus on a large corpus for our experiments. However, this is one direction of our future work. *Semantic indexing* is the process which extracts concepts within documents or queries in natural language [9]. The aim is to find the most representative concepts for documents and queries. We use a program made in our lab : RIIO [2], which is based on the selection of synsets from WordNet. Thus, there is no human intervention in the process. We use a *similarity* function based on [1], because it has good properties and results. The *propagation* functions used are of the form $f_{1,l_2,v}$ (see figure 1)

We compare our *image based method* with two others that are classicaly used in the context of a shared ontology. In the *cosine* based method, relevance is defined by the cosine between the query and the document vectors. In the *rough*

[1] It was collected between 1957 and 1968 by Cyril W. Cleverdon. The documents are abstracts of research papers.

expansion method, the effects of propagating weights from different concepts are mixed in a single vector. Relevance is obtained using the cosine. This method avoids some silence, but often generates too much noise, without any highly accurate sense disambiguation [10]. In the context of a shared ontology, our method shows *i)* better results than rough expansion and *ii)* results that are comparable with the cosine ones (a 2% increase of recall and precison). In the context of two different ontologies, the cosine based method is applied by the document provider p_2 in space $\mathcal{C}_{\Omega_2}$: in the query, only the shared concepts are considered. Rough expansion is done at the query intiator p_1's and the cosine is calculted at p_2's.

Because the manipulation of two different ontologies is a heavy process, we decided to *simulate* that use. Wordnet is used by both p_1 (to express its queries) and p_2 (to index its documents). However, in the result of the ontology mapping, we randomdly remove a given percentage of the shared concepts (from 10% to 90%). This amounts to simulate two peers that use the same ontology but are not aware of it. Of course, this eases interpretation. In particular, taking the lowest common ancestor of the shared concepts of $\overrightarrow{sed}_c$ to find the corresponding concept of central concept c gives good results most of the time.

Figure 5 shows the results obtained in average for the first twelve queries of the testing corpus. The reference method is the cosine one when no concept is removed, which gives a given reference precision and recall. Then, for each method and each percentage of removed concepts, we compute the ratio of the precision obtained (respectively recall) by the reference precision. When the percentage of randomly removed concepts increases, precision (figure 5 (a)) and recall (Figure 5 (b)) decrease i.e. the results are less and less relevant. However, our image and interpretation based solution shows much better results. When the percentage of removed concepts is under 70%, we still get 80% or more of the answers obtained in the reference case.

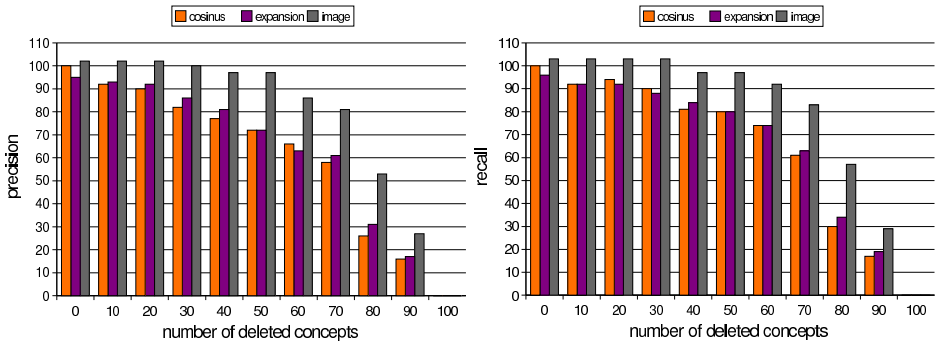

Fig. 5. Evolution of (a) precision and (b) recall in function of the percentage of concepts randomly removed from the set of shared concepts.

6 Conclusion

The main contribution of this paper is a solution which improves information exchange between two peers that use different ontologies. Our solution uses semantic vectors to represent documents and queries. It only requires the peers to share some concepts and uses the unshared concepts to find additional relevant documents. To the best of our knowledge, the problem has never been addressed before and our approach is a first, encouraging solution.

When performing query expansion, the query initiator makes more precise the concepts of the query by associating an expansion to each of them (SED). The expansion depends on the initiator's characteristics: ontology, similarity, propagation function. Interpretation by the document provider is not easy because the peers do not share the vector space. Given its own ontology and similarity function, it first finds out a corresponding concept for the central concept of each SED, and then interprets the whole SED. The interpreted SEDs are used to compute an image of the documents and their relevance. This is only possible because the central concepts are expanded separately. The results of our preliminary experiments show that our approach significantly improves information exchange, finding up to 90% of the documents that would be found if all the concepts were shared.

As future work, we plan to conduct additional testing in different contexts to verify the robustness of our approach. Another aspect that we want to consider carefully is complexity to improve efficiency. Finally, we plan a theoretical study of the impact of interpretation when several peers are involved

References

1. A. Bidault, C. Froidevaux, and B. Safar. Repairing queries in a mediator approach. In *ECAI*, 2000.
2. E. Desmontils and C. Jacquin. *The Emerging Semantic Web*, chapter Indexing a web site with a terminology oriented ontology. 2002.
3. J. Euzenat and P. Shvaiko. *Ontology matching*. Springer-Verlag, 2007.
4. A. Gómez-Pérez, M. Fernández, and O. Corcho. *Ontological Engineering*. Springer-Verlag, London, 2004.
5. Z. G. Ives, A. Y. Halevy, P. Mork, and I. Tatarinov. Piazza: mediation and integration infrastructure for semantic web data. *Journal of Web Semantics*, 2003.
6. J.-Y. Nie and F. Jin. Integrating logical operators in query expansion invector space model. In *SIGIR workshop on Mathematical and Formal methods in Information Retrieval*, 2002.
7. P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *IJCAI*, 1995.
8. M.-C. Rousset. Somewhere: a scalable p2p infrastructure for querying distributed ontologies. In *CoopIS/DOA/ODBASE*, 2006.
9. M. Sanderson. Retrieving with good sense. *Information Retrieval*, 2000.
10. E. M. Voorhees. Query expansion using lexical-semantic relations. In *SIGIR*, Dublin, 1994.
11. W. Woods. Conceptual indexing: A better way to organize knowledge. Technical report, Sun Microsystems Laboratories, 1997.