Query interpretation to help peers understand each others in semantically heterogeneous systems

Anthony Ventresque, Sylvie Cazalens, Philippe Lamarre, Patrick Valduriez

Improving Interoperability Using Query Interpretation in Semantic Vector Spaces

Anthony Ventresque* Sylvie Cazalens* Philippe Lamarre* Patrick Valduriez‡

Atlas group, INRIA and LINA, Université de Nantes
2, rue de la Houssinière, 44322 Nantes - FRANCE
E-mail: Firstname.Lastname@{univ-nantes.fr*, inria.fr‡}

Abstract

In semantic web applications where query initiators and information providers do not necessarily share the same ontology, semantic interoperability generally relies on ontology matching or schema mappings. Information exchange is then not only enabled by the established correspondences (the "shared" parts of the ontologies) but, in some sense, limited to them. Then, how the "unshared" parts can also contribute to and improve information exchange ? In this paper, we address this question by considering a 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. Our experiments show an important improvement of retrieval relevance when concepts of documents and queries are not shared. Even if the concepts of the initial query are not shared by the document provider, our method still ensures 90% of the precision and recall obtained when the concepts are shared.

1. Introduction

In semantic web applications where query initiators and information providers do not necessarily share the same ontology, semantic interoperability generally relies on ontology matching or schema mappings. Several works in this domain focus on what (*i.e.* the concepts and relations) the peers share [9, 18]. This is quite important because, obviously if nothing is shared between the ontologies of two peers, there is a little chance for them to understand the meaning of the information exchanged. 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. The question is then to know whether the unshared parts can still be useful for information exchange.

In this paper, we focus on semantic interoperability and information exchange between a query initiator p_1 and a document provider p_2, which use different ontologies but share some common concepts. 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 shared.* This problem is very important because in semantic web applications with high numbers of participants, the ontology (or ontologies) is rarely entirely shared. Most often, participants agree on some part of a reference ontology to exchange information and internally, keep working with their own ontology [18, 21].

We represent documents and queries by *semantic vectors* [24], a model based on the vector space model [1] using concepts instead of terms. Although there exist other, richer representations

(conceptual graphs for example), semantic vectors are a common way to represent unstructured documents in information retrieval. 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 ontologies, 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 for the query initiator that would not be found by only using the matching concepts in the query and the documents. Although the basic idea of query expansion and interpretation is simple, query interpretation is very difficult because it requires to precisely weight additional concepts given some weighted shared ones, while the whole space (i.e. the ontology) and similarity measures change.

In this context, our contributions are the following. First, we propose a specific query expansion method. Its property is to keep separate the results of the propagation from each central concept of the query, thus limiting the noise due to inaccurate expansion. Second, given this expansion, we define the relevance of a document. Its main, original characteristic is to require the document vector to be requalified with respect to the expanded query, the result being called *image* of the document. Third, a main contribution is the definition of query interpretation which enables the expanded query to be expressed with respect to the provider's ontology. Finally, we provide two series of experiments. In the first one, with a single shared ontology, we verify that our query expansion and relevance calculus show results which are comparable to the standard query expansion ones [23, 13]. In the second experiment, we introduce unshared concepts and we still find up 90% of the documents that would be selected if all the central concepts were shared. To the best of our knowledge, the problem of improving information exchange by using the unshared concepts of different ontologies has not been addressed before. Our proposal is a first, encouraging solution.

This paper is organized as follows. Section 2 gives preliminary definitions. Section 3 presents our query expansion method and the image based relevance of a document. For simplicity, we assume a context of shared ontology. This assumption is relaxed after in Section 4, where we consider the case where the query initiator and the document provider use different ontologies and present the query interpretation. Section 5 discusses the experiments and their results. The two last sections are respectively devoted to related work and conclusion.

2. Preliminary Definitions

2.1. Semantic vectors

In the vector space model [1], documents and queries in natural language are represented as vectors of keywords (terms). If there are n keywords, each document is represented by a vector in the $n-$dimensional space. Relevance of a document can then been calculated by comparing the deviation of angles between the document vector and the original query vector. An approach based on *semantic* vectors [24, 11] uses the same kind of multi-dimensional linear space except that it no longer considers keywords but *concepts* of an ontology: 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. Notice that, although the experiments are conducted with text documents and natural language queries, the approach is very general and can be used whenever queries and doc-

uments can be represented by semantic vectors.

Obviously, the approach relies on the use of an ontology. It is in no case the point of the paper to discuss what an ontology is [7]. We simply define an ontology as a set of concepts together with a set of relations between those concepts. In this paper, the illustrations and the experiments use an ontology where the only relation considered is an is-a relation (specialization link). However, this does not restrict the generality of our relevance calculus. Indeed, the presence of several relations only affects the definition of the similarity of a concept with respect to another. In the rest of the paper, we assume the existence of an ontology Ω, $\mathcal{C}_\Omega$ being its set of concepts.

Definition 1 *A* semantic vector $\overrightarrow{v_\Omega}$ *is an application defined on the set of concepts* $\mathcal{C}_\Omega$ *of the ontology:*

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

Without loss of generality, we consider real values in the interval $[0,1]$. Reference to the ontology will be omitted whenever there is no ambiguity. Usually, the concepts of the semantic vector (i.e. those of the ontology) are also called the dimensions of the vector. For example, let us consider an ontology where $\mathcal{C}_\Omega = \{c_1, c_2, c_3\}$, and a document characterized across these three dimensions by the semantic vector $\overrightarrow{v}$ with $\overrightarrow{v}[c_1] = 0.2$, $\overrightarrow{v}[c_2] = 0.7, \overrightarrow{v}[c_3] = 0$. The meaning is that the document is related to concept c_1 (but probably c_1 is not central), strongly related to concept c_2, and not related to concept c_3. The vector has two non-null dimensions, which we will also call *weighted concepts*. A popular way to compute the relevance of a document is to use the cosine-based proximity of the document and query vectors (respectively $|\overrightarrow{d}|$ and $|\overrightarrow{q}|$) in the space, as in the following formula where $|\overrightarrow{d}|$ represents the norm of $\overrightarrow{d}$:

$$\cos(\overrightarrow{d}, \overrightarrow{q}) = \frac{\overrightarrow{d} \cdot \overrightarrow{q}}{|\overrightarrow{d}| \times |\overrightarrow{q}|}$$

2.2. Similarity and propagation

Query expansion is generally used to find additional relevant documents. In a semantic vectors setting, the problem can be expressed as follows: given the weighted concepts of a semantic vector, how should we weight those initially unweighted concepts which seem linked to the weighted ones ? In other words, given some weighted concept c_1, called *central concept*, how should we propagate its weight on the other linked concepts ? An intuitive solution is to consider that the more *similar* to c_1 an unweighted concept is, the more the *propagation* should weight it. Similarity of some concept wrt some central concept c_1 can be represented as a function which values each concept of the ontology according to its similarity degree with c_1. Instead of considering the function, we consider the induced ordering of the concepts of $\mathcal{C}_\Omega$.

Definition 2 *Let* c *be a concept of* Ω. 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* $\mathcal{C}_\Omega$.

Definition 3 (Propagation function) *Let* c *be a concept of* Ω *valued by* v*; and let* sim_c *be a similarity function. A function* $\mathcal{P}f_c$: $[0..1] \mapsto [0..1]$ $sim_c(c') \rightarrow \mathcal{P}f_c(sim_c(c'))$ *is a* propagation function from c *iff*

- $\mathcal{P}f_c(sim_c(c)) = v$*, and*
- $\forall c_k, c_l \in \mathcal{C}_\Omega \; sim_c(c_k) \leq sim_c(c_l) \Rightarrow \mathcal{P}f_c(sim_c(c_k)) \leq \mathcal{P}f_c(sim_c(c_l))$

2.3. Example of propagation function (fig 1)

In practice, we have tested several types of functions to be used in our query expansion method. A class of propagation functions from c which works fine is inspired by the membership functions used in fuzzy logic [26]. It is defined by three parameters v (weight of the central concept), l_1 (length of the interval where concepts have the same weight : v) and l_2 (length of the interval where concepts have non zero weight) such that:

$$\mathcal{P}f_c(x) = f_{v,l_1,l_2}(x) = \begin{cases} v & if\ x \geq l_1 \\ \frac{v}{l_1 - l_2}x + \frac{l_2 \times v}{l_1 - l_2} & if\ l_1 > x > l_2 \\ 0 & if\ l_2 \geq x \end{cases}$$

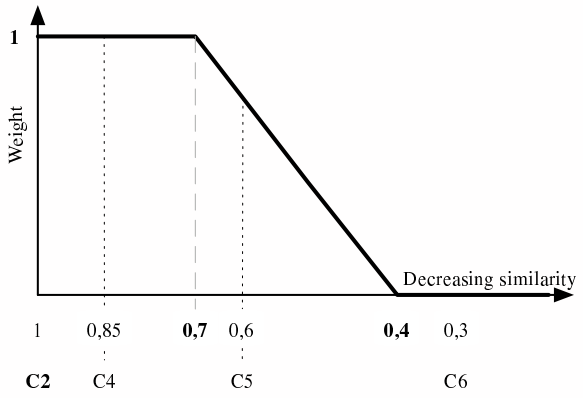


Figure 1: Example of a propagation function $f_{1,0.7,0.4}$ with central concept c_2 .

3. Query expansion and Image based relevance

In this section we present our method to compute the relevance of a document wrt a query. For the sake of simplicity, we assume that the query initiator and the document provider use the same ontology. Notice however, they can still differ on the similarity measures and the propagation functions. Our method has two main steps. The first one is *query expansion*, which general aim is to weight relevant concepts that are not initially shared by the query and document vectors. The second one is a *relevance* computation, which uses the *image of the document vector*.

3.1. Query expansion

In the case of a single ontology, query expansion can be carried out by the query initiator or by the document provider. To our view, the query initiator side might be better because it knows better what it is looking for. However, processing expansion at the provider's, would just amount supporting the provider's point of view on similarity and propagation rather than the query initiator's one.

To our knowledge, most propagation methods propagate the weight of each weighted concept in *the same vector*, thus directly adding the expanded terms in the original vector. When a concept is involved in several propagations conducted from different central concepts, a kind of aggregation function (like for example the maximum) is used. We call this kind of method "rough" propagation. Although its results are not bad, such a propagation figures out some drawbacks among which a possible unbalance of the relative importance of the initial concepts [16]. For example, assume that the initial query vector is $\overrightarrow{q}[c_1] = 0.5$ and $\overrightarrow{q}[c_2] = 0.5$. Then propagation from c_2 also weights concepts c_3, c_4 and c_5, respectively with weights 0.3, 0.3 and 0.2; but propagation from c_1 weights no other concept. In that case, one could say that propagation attributes more importance to the initial concept c_2. This is a reason why our choice is to keep separate the results of the propagation from different concepts of the query. In addition, as we shall see in section 4, this choice eases query interpretation, which would be more difficult if all the effects of propagations would be mixed in the initial query vector.

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. One may consider all the weighted concepts in $\overrightarrow{q}$ (as we did in our experiments) or just those which weight is above some threshold. To keep separate the effects of different propagations, each central concept of $\mathcal{C}_{\overrightarrow{q}}$ is *semantically enriched* by propagation, in a separate vector.

Definition 4 (Semantically Enriched Dimension) *Let $\overrightarrow{q}$ be a query vector and let c be a concept in $\mathcal{C}_{\overrightarrow{q}}$. A semantic vector $\overrightarrow{sed}_c$ is a* semantically enriched dimension, *iff* $\forall c' \in \mathcal{C}_\Omega, \overrightarrow{sed}_c[c'] \leq \overrightarrow{sed}_c[c]$.

Definition 5 (Expansion of a query) *Let $\overrightarrow{q}$ be a query vector. An* expansion *of $\overrightarrow{q}$, noted $\mathcal{E}_{\overrightarrow{q}}$ is a set defined by:*
$\mathcal{E}_{\overrightarrow{q}} = \{\overrightarrow{sed}_c : c \in \mathcal{C}_{\overrightarrow{q}},\ \forall c' \in \mathcal{C}_\Omega, \overrightarrow{sed}_c[c'] = \mathcal{P}f_c(c')\}$

Figure 2 illustrates the expansion of a query $\overrightarrow{q}$ with two weighted concepts c_4 and c_7. It contains two semantically enriched dimensions. In dimension $\overrightarrow{sed}_{c_7}$, concept c_7 has the same value as in the query. The weight of c_7 has been propagated on c_3, c_{11} and c_6 according to their similarity with c_7. The other dimension is obtained from c_4 in the same way.

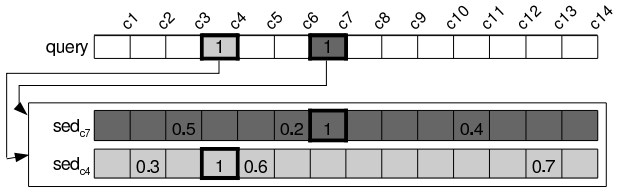


Figure 2: A query expansion composed of 2 semantically enriched dimensions.

3.2. Image of a Document and Relevance

Intuitively, our aim is to better compare the query and document representations. A way to do so, is to be able to characterize the document wrt each central concept c (dimension) of the query, as far as it has concepts related to c, in particular even if c is not initially weighted in $\overrightarrow{d}$. In other words, if concept c' is linked to concept c, c' being weighted in $\overrightarrow{d}$, although it isn't in $\overrightarrow{q}$, the corresponding dimensions in the space should not be considered as independent. This idea is implemented in the definition of the image of the document.

Given a SED $\overrightarrow{sed}_c$, we aim at valuating c in the image of the document $\overrightarrow{d}$ according to the relevance of $\overrightarrow{d}$ to $\overrightarrow{sed}_c$. To evaluate the impact of $\overrightarrow{sed}_c$ on $\overrightarrow{d}$ we consider the product of the respective values of each concept in $\overrightarrow{sed}_c$ and $\overrightarrow{d}$. Intuitively all the concepts of the document which are linked to c through $\overrightarrow{sed}_c$ have a nonnull value. The image of $\overrightarrow{d}$ keeps track of the best value attributed to one of the linked concepts if it is better than $\overrightarrow{d}[\ c\ ]$, which is the initial value of c . This process is repeated for each SED of the query. Algorithm 1 details the calculus of the image of document $\overrightarrow{d}$, noted $\overrightarrow{i}_d$.

This algorithm ensures that all the central concepts of the initial query vector are also weighted in the image of the document as far as the document is related to them. With respect to the query, the image of the document is more accurate because it somewhat enforces the documents characterization over each dimension of the query. However, as can be noticed in the algorithm, in the image, we keep unchanged the weights of the concepts which are not linked to any concept of the query (i.e. which are not weighted in any SED). This has some importance when the document has many unrelated concepts, for example if it is very general. In that case, the norm of the vector gets higher (and consequently, its relevance lower).

Algorithm 1: Image of a document wrt a query.

input : a semantic vector $\overrightarrow{d}$ on an ontology Ω; an expanded query $\mathcal{E}_{\overrightarrow{q}}$
output: a semantic vector $\overrightarrow{i}_d$, image of $\overrightarrow{d}$.
begin
  forall $c \in \mathcal{C}_{\overrightarrow{q}}$ **do**
    forall $c' : \overrightarrow{sed}_c[c'] \neq 0$ **do**
      $\overrightarrow{i}_d[c] \leftarrow$
      $\max(\overrightarrow{d}[c'] \times \overrightarrow{sed}_c[c'], \overrightarrow{i}_d[c])$;
  forall $c \notin \mathcal{C}_{\overrightarrow{q}}$ **do**
    if $\exists c' \in \mathcal{C}_{\overrightarrow{q}} : \overrightarrow{sed}_{c'}[c] \neq 0$ **then**
      $\overrightarrow{i}_d[c] \leftarrow 0$
    else
      $\overrightarrow{i}_d[c] \leftarrow \overrightarrow{d}[c]$;
  return $\overrightarrow{i}_d$;
end

The example of figure 3 illustrates the calculus of the image of a document. Each SED of the expanded query is combined with the semantic vector of the document. Let us consider $\overrightarrow{sed}_{c_4}$. In the document, the weight of c_4 is null. However, the semantically enriched dimension related to c_4 weights other concepts. In particular, we have $\overrightarrow{sed}_{c_4}[\ c_2\ ] = 0.3$. As $\overrightarrow{d}[\ c_2\ ] = 1$, the resulting product is 0.3. Because this value improves $\overrightarrow{d}[\ c_4\ ]$ (which is null) we keep it in the image of

the document. Hence, in the image, we can express that the document is related to concept c_4 of the query, even if it wasn't the case initially. Notice that concepts c_1 and c_9, which are linked to no SED keep their initial values.

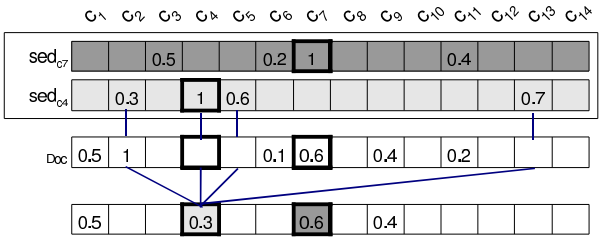


Figure 3: Obtaining the image of a document.

The core of this type of expansion lies here. Creating and using a SED amounts to unfold additional nonnull dimensions to help recovering additional information about the semantics of the document, and then fold them again to evaluate the relevance of the document.

3.2.1. Relevance of a document

We adopt a global measure of relevance using the image of the document and the cosine.

Definition 6 *Let $\overrightarrow{d}$ and $\overrightarrow{q}$ be respectively a document and query vector and let $\overrightarrow{i}_d$ be the image of $\overrightarrow{d}$. Then the relevance of $\overrightarrow{d}$ wrt $\overrightarrow{q}$ is:*

$$Relevance(\overrightarrow{d}, \overrightarrow{q}) = \frac{\overrightarrow{i}_d \cdot \overrightarrow{q}}{|\overrightarrow{i}_d| \times |\overrightarrow{q}|}$$

Considering the image enables to take into account the documents that have concepts linked to those of the query. Using a cosine, and thus the norm of the vectors, attributes a lower importance to the documents with an important norm, which are often very general.

3.2.2. Complexity of the process

Concerning the complexity of the expansion, the theorical complexity is always $O(n)$ (where n is the number of concepts of the ontology), because of the small number of dimensions in the query. Theoretically, in the worst case, computation of the image is $O(n^2)$. However, in practice, the complexity remains reasonable because the number of concepts of the query remains very small (5 in average according to our experiments) compared to the number of concepts in the ontology. Relevance evaluation itself is $O(n)$.

4. Relevance in the context of unshared concepts

In this section, we assume that the query initiator and the document provider do not use the same ontology. However, if they wouldn't be able to establish any correspondence between some of their own concepts, interoperability would be impossible. Thus we consider that they *share* some common concepts, meaning that each of them regularly (although may be not often) computes an ontology matching algorithm which provides a set of correspondences (equivalences) between those concepts. This is a minimal requirement for interoperability.

However, this doesn't mean that the unshared concepts are of no use to evaluate the relevance of a document wrt a query. The problem is how to take them into account. To do so, we keep on with the philosophy adopted in section 3: we still use a query expansion at the query initiator's side and the calculus of the image of the document at the provider's side. Things are complicated by the fact that the query initiator and the document provider do not use the same vector space. This is why we introduce a query *interpretation* step at the provider's side. The interpreted query is used to compute the image of the document.

4.1. Computing Relevance: Overview

As shown in Figure 4, the query initiator, denoted by p_1, works within the context of ontology Ω_1, while the document provider, noted p_2, works with ontology Ω_2. Through its semantic indexing module, the query initiator (respectively the document provider) produces the query vector (respectively the document vector), which is expressed on Ω_1 (respectively Ω_2). Both p_1 and p_2 also have their own way of computing both the

similarity and the propagation.

We assume that the query initiator and the document provider *share* some common concepts, meaning that each of them regularly, although may be not often, runs an ontology matching algorithm. Ontology matching results in an *alignment* between two ontologies, which is composed of a (non empty) set of correspondences with some cardinality and, possibly some meta-data [4]. A *correspondence* establishes a relation (equivalence, subsumption, disjointness...) between some entities (in our case, concepts), with some confidence measure. Each correspondence has an identifier. In this paper, we only consider the equivalence relation between concepts and those couples of equivalent concepts of which confidence measure is above some threshold. We call them the *shared* concepts. For simplicity, when there is an equivalence, we make no distinction 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. Hence, the set of shared concepts is denoted by $\mathcal{C}_{\Omega_1} \cap \mathcal{C}_{\Omega_2}$.

Given these assumptions, computing relevance requires the following steps :

Query Expansion. It remains unchanged. The query initiator 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 of a SED $\overrightarrow{sed}_c$ is decomposed in two problems, which we address in the next subsections:

- The first problem is to find a concept in $\mathcal{C}_{\Omega_2}$ that corresponds to c, noted $\tilde{c}$. This is difficult when the central concept is not shared. In this case, we use the weights of the shared concepts to guide the search. Of course, this is only a "contextual" correspondence as opposed to one that would be obtained through matching.

- The second problem is to attribute weights to shared and unshared concepts of $\mathcal{C}_{\Omega_2}$ which are linked to $\overrightarrow{sed}_c$. This amounts to interpret the SED.

Image of the Document and Cosine Computation. They remain unchanged. Provider p_2 computes the image of its documents wrt. the interpreted SEDs and then, their cosine based relevance wrt. the interpreted query, no matter the ontology used by p_1.

In the following, we describe the steps involved in the interpretation of a given SED.

4.2. Finding a Corresponding Concept

The interpretation of a given SED $\overrightarrow{sed}_c$ leads to a major problem: finding a concept in $\mathcal{C}_{\Omega_2}$ which corresponds to the central concept c. This corresponding concept is noted $\tilde{c}$ and will play the role of the central concept in the interpretation of $\overrightarrow{sed}_c$, noted $\overrightarrow{sed}_{\tilde{c}}$. If c is shared, we just keep it as the central concept of the interpreted SED. When c is not shared we have to find a concept which seems to best respect the "flavor" of the initial SED.

Theoretically, all the concepts of $\mathcal{C}_{\Omega_2}$ should be considered. Several criterias can apply to choose one which seems to best correspond. We propose to define the notion of *interpretation function* which is relative to a SED $\overrightarrow{sed}_c$ and a candidate concept $\tilde{c}$ and which assigns a weight to each value of silmilarity wrt. $\tilde{c}$. Definition 7 consists of four points. The first one requires the interpretation function to assign the value of $\overrightarrow{sed}_c[c]$ to the similarity value 1, which corresponds to $\tilde{c}$. In the second point, we use the weights assigned by $\overrightarrow{sed}_c$ to the shared concepts (c_1, c_2, c_3 and c_6 in figure 5) and the ranking of concepts in function of $sim_{\tilde{c}}$. However, there might be several shared concepts that have the same similarity value wrt. $\tilde{c}$, but have a different weight according to $\overrightarrow{sed}_c$. Thus, we require function $f_i^{\overrightarrow{sed}_c,\tilde{c}}$ to assign the minimum of these values to the corresponding similarity value. This is a pessimistic choice and we could either take the maximum or a combination of these weights. As for the third point, let us call c_{min}, the shared concept with the lowest similarity value (c_6 in Figure 5 (a) and c_3 in Figure 5

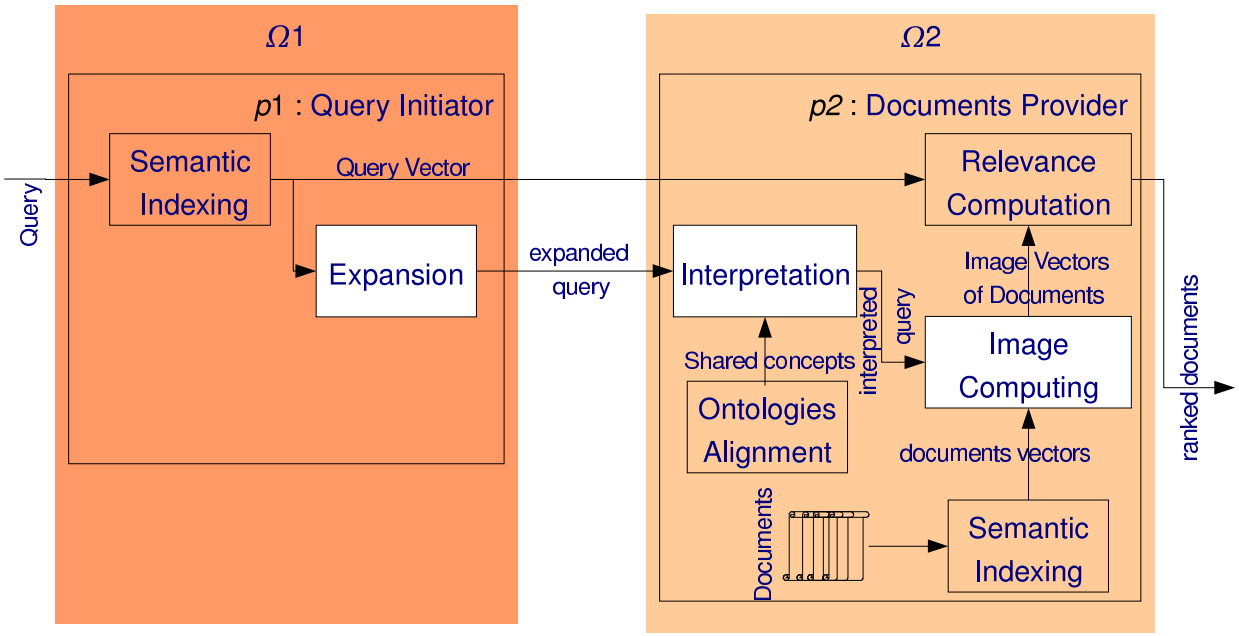


Figure 4: Overview of relevance computation

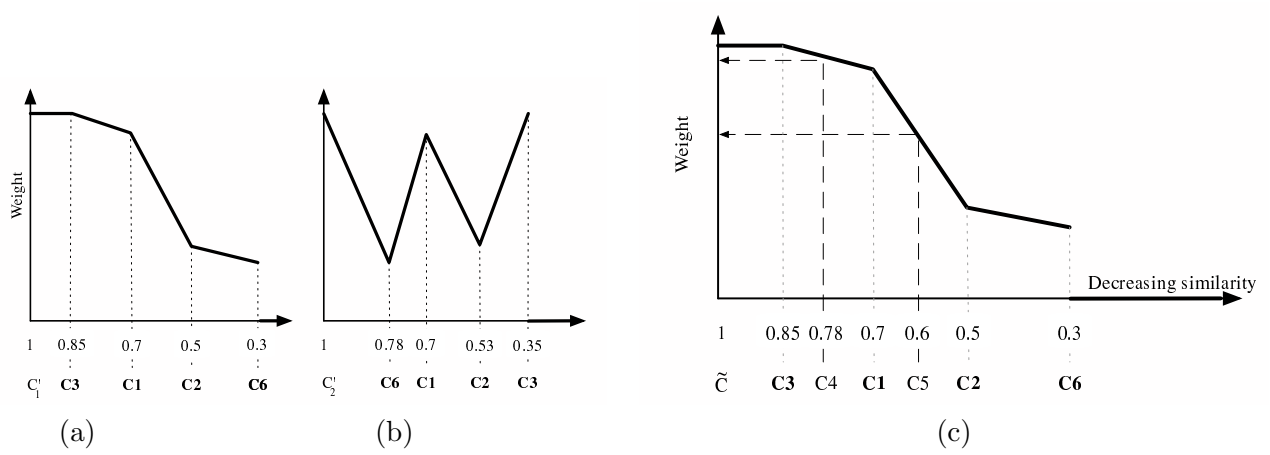
(b)). We consider that we have not enough information to weight the similarity values lower than $sim_{\tilde{c}}(c_{min})$. Thus we assign them the zero value. The fourth point is just a mathematical expression which ensures that the segments of the affine function are only those defined by the previous points.

Definition 7 (Interpretation function) *Given a SED* $\overrightarrow{sed}_c$ *and a concept* $\tilde{c}$, $f_i^{\overrightarrow{sed}_c,\tilde{c}} : [0..1] \rightarrow [0..1]$, *noted* f_i *if no ambiguity, is an* interpretation function *iff it is a piecewise affine function and:*

- $f_i(1) = \overrightarrow{sed}_c[c]$;
- $\forall c' \in \mathcal{C}_{\Omega_1} \cap \mathcal{C}_{\Omega_2}$, $f_i(sim_{\tilde{c}}(c')) = \min_{\substack{c'' \in \mathcal{C}_{\Omega_1} \cap \mathcal{C}_{\Omega_2} \\ sim_{\tilde{c}}(c') = sim_{\tilde{c}}(c'')}} (\overrightarrow{sed}_c[c''])$;
- $\forall x \in [0..1]$, $x < sim_{\tilde{c}}(c_{min}) \Rightarrow f_i(x) = 0$;
- $Seg = \|\{x : \exists c' \in \mathcal{C}_{\Omega_1} \cap \mathcal{C}_{\Omega_2}, c' \neq \tilde{c} \text{ and } sim_{\tilde{c}}(c') = x\}\| + 1$ *where* Seg *is the number of segments of* f_i.

Intuitively, the criterias for choosing a corresponding concept among all the possible concepts can be expressed in terms of the properties of the piecewise affine function f_i. Of course, there are as many different function f_i as candidate concepts. But the general idea is to choose the function f_i wich resembles the more to a propagation function. Let us consider the example of Figure 5 (a) and (b) where c_1, c_2, c_3 and c_6 are shared. The function in Figure 5 (a) is obtained considering c'_1 as the corresponding concept (and thus ranking the other concepts in function of their similarity with $c'1$). The function in Figure 5 (b) is obtained similarly, considering c'_2. Having to choose between c'_1 and c'_2 we would prefer c'_1 because function $f_i^{\overrightarrow{sed}_c,c'_1}$ is monotonically decreasing whereas $f_i^{\overrightarrow{sed}_c,c'_2}$ shows a higher "disorder" wrt. the general curve of a propagation function.

Several characteristics of the interpretation function can be considered to evaluate "disorder". For example, one could choose the function which minimizes the number of local minima (thus minimizing the number of times the sign of the derivated function changes). Another example is to choose the function which minimizes the variations of weight between local minima and their next local maximum (thus penalizing the functions which do not decrease monotonically). A third could combine these criteria.



(a) (b) (c)

Figure 5: Two steps of the interpretation of a SED : (a) f_i for candidate concept c'_1, (b) f_i for candidate concept c'_2 and (c) weighting the unshared concepts.

4.3. Interpreting a SED

We define the interpretation of a given SED $\overrightarrow{sed}_c$ as another SED, with central concept $\tilde{c}$ which has been computed at the previous step. We keep their original weight to all the shared concepts. The unshared concepts are weighted using an interpretation function as defined above.

Definition 8 (Interpretation of a SED) *Let $\overrightarrow{sed}_c$ be a SED on $\mathcal{C}_{\Omega_1}$ and let $\tilde{c}$ be the concept corresponding to c in $\mathcal{C}_{\Omega_2}$. Let $sim_{\tilde{c}}$ be a similarity function and let $f_i^{\overrightarrow{sed}_c,\tilde{c}}$, noted f_i, be an interpretation function. Then SED $\overrightarrow{sed}_{\tilde{c}}$ is an interpretation of $\overrightarrow{sed}_c$ iff:*

- $\overrightarrow{sed}_{\tilde{c}}[\tilde{c}] = f_i(1)$;
- $\forall c' \in \mathcal{C}_{\Omega_1} \cap \mathcal{C}_{\Omega_2},\ \overrightarrow{sed}_{\tilde{c}}[c'] = \overrightarrow{sed}_c[c']$;
- $\forall c' \in \mathcal{C}_{\Omega_2} \setminus \mathcal{C}_{\Omega_1},\ \overrightarrow{sed}_{\tilde{c}}[c'] = f_i(sim_{\tilde{c}}(c'))$;

Figure 5 (c) illustrates this definition. Document provider $p2$ ranks its own concepts in function of $sim_{\tilde{c}}$. Among these concepts, some are shared ones for which the initial SED $\overrightarrow{sed}_c$ provides a given weight. This is the case for c_1, c_2, c_3 and c_6 which are in bold face in the figure. The unshared concepts are assigned the weight they obtain by function f_i (through their similarity to $\tilde{c}$). This is illustrated for concepts c_4 and c_5 by a dotted arrow.

5. Experimental Validation

In this section, we use our approach based on *image based relevance* to find documents which are the most relevant to given queries. We compare our results with those obtained by the *cosine based method* and the *rough propagation method*. In the former method, relevance is defined by the cosine between the query and document vectors. In the latter, the effects of propagating weights from different concepts are mixed in a single vector; then relevance is obtained using the cosine.

5.1. General Setup for the Experiments

We use the Cranfield corpus, a testing corpus consisting of 1400 documents and 225 queries in natural language, all related to aeronautical engineering. For each query, each document is scored by humans as relevant or not relevant (boolean relevance). Our ontology is lightweight, in the meaning of [7], *i.e.* an ontology composed of a taxonomy of concepts : WordNet [5]. In Information Retrieval, there was a debate whether WordNet is suitable for experimentation (see the discussion in [23]). However, more recent works show that it is possible to use WordNet, and sometimes other resources, and still get good results [8]. Semantic indexing [19] is the process which can compute the semantic vectors from documents or queries in natural language. The aim is to find the most representative concepts for documents or queries. We

use a program made in our lab : RIIO [3], which is based on the selection of synsets from WordNet. Although it is not the best indexing module, one of its advantages is that there is no human intervention in the process. The semantic similarity function we use is that of [2], because it has good properties and results which are discussed in Section 6. We slightly modified that function due to normalization considerations. Following the framework of membership functions presented in Section 2.3 we can define many propagation functions. We tested three different types of functions : "square" (of type f_{v,l_1,l_1}), "sloppy" (of type $f_{v,1,l_2}$), or hybrid (of type f_{v,l_1,l_2} with $l_1 = 2 \times l_2$). Our experiments show no important difference, but sloppy propagation has slightly better results. So we use only this propagation function, adding ten concepts in average for a given central concept.

In order to evaluate whether our solution is robust, we would need ontologies which agree on different percentages of concepts : 90%, 80%, 70%, ..., 10%. This is very difficult to obtain. We could build artificial ontologies, but this would force us to give up the experiments on a real corpus. Thus, we decided to stick to WordNet and simulate semantic heterogeneity. Both the query initiator and the provider use WordNet, but we make so that they are not able to understand each other on some concepts (a given percentage of them). To do so, we remove some mappings between the two ontologies. Thus it simulates the case where the query intiator and the document provider use the same ontology but are not aware of it. It is then no more possible to compare queries and documents on those concepts. The aim is to evaluate how the answers to queries expressed with removed matchings, change. Note that the case with no removed matching reduces to a single ontology.

In a first experiment, we progressively reduce the number of mappings, thus increasing the percentage of removed mappings (10%, 20%, ... until 90%). The progressive reduction in their common knowledge is done randomly. In a second experiment, we remove the mappings concerning the central concepts of the queries in the ontology of the document manager. This is now an intentional removing, which is the worst case for most of the techniques in IR : removing only the elements that match. For both experiments, we take into account the results obtained with the 225 queries of the corpus.

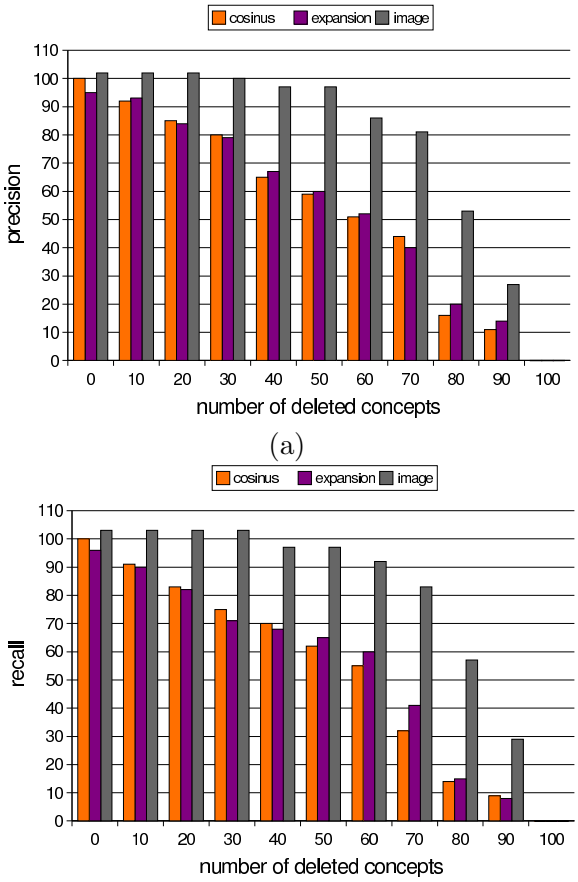


Figure 6: Evolution of (a) precision and (b) recall in function of the random removal percentage of mappings.

5.2. Results

Figure 6 shows the results obtained in average for the all 225 queries of the testing corpus. The reference method is the cosine one when no matching is removed, which gives a given reference precision and recall. Then, for each method and each percentage of removed matching, we compute the ratio of the precision obtained (respectively recall) by the reference precision. When the percentage

of randomly removed matchings increases, precision (Figure 6 (a)) and recall (Figure 6 (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 matchings is under 70%, we still get 80% or more of the answers obtained in the reference case.

In the second experiment, we consider that the document manager does not understand (*i.e.* share with the query initiator) the central concepts of the query (see Figure 7). With the cosine method, there is no more matching between concepts in queries and concepts in documents. Thus no relevant document could be retrieved. With the query expansion, some of the added concepts in the query allow to match with concepts in documents that are close to the central concepts of the query. This leads to precision and recall at almost 10%. Our image-based retrieving method has more than 90% of precision and recall in the retrieval. This is also an important result. Obviously, as we have the same ontology and the same similarity function, the interpretation can retrieve most of the central concepts of the query. But the case presented here is hard for most of the classical techniques (concepts of the query unshared) and we obtain a very important improvement.

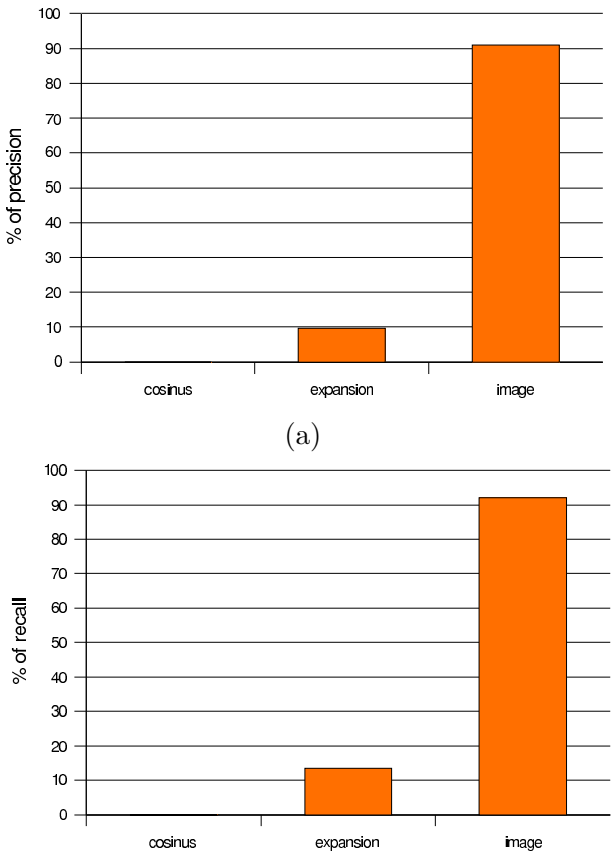


Figure 7: Precision (a) and recall (b) when the central concepts of the query are unshared.

6. Related Work

The similarity that we use in our experiments is the result of a thorough study of the properties of different similarity measures. We looked for a similarity which is not a distance (does not satisfy similarity nor triangle inequality), based on the result of [22]. Hence we use one classical benchmark of this domain : the work of Miller and Charles [15] on the human assessments of similarity between concepts. Thirty eight students were asked to mark how similar thirty couples of concepts were. We have implemented four similarity measures: [25, 20, 12, 2], respectively noted *Wu and P.*, *Seco*, *Lin* and *Bidault* in table 1. Correlation is the ratio between those measures on the human results. The results show that only Bidault's measure does not meet symmetry nor triangle inequality. Moreover, it obtains a slightly better correlation. Hence, it was preferred to rank the concepts according to their (dis)similarity with a central concept.

	Wu & P.	Seco	Lin	Bidault
symmetry	yes	yes	yes	**no**
triangle inequality	no	no	no	no
correlation	0.74	0.77	0.80	0.82

Table 1: Comparison of similarity measures.

The idea of query expansion is shared by several fields. It was already used in the late 1980's in Cooperative Answering Systems [6]. Some of the suggested techniques expanded SQL queries

considering a taxonomy. In this paper, we do not consider SQL queries, and we use more recent results about ontologies and their interoperability. Expansion of query vectors is used for instance in [17, 23]. However, this expansion produces a single semantic vector only. This amounts to mix the effects of the propagations from different concepts of the query. Although this method avoids some silence, it often generates too much noise, without any highly accurate sense disambiguation [23]. Consequently, the results can be worse than in the classical vector space model [1]. Our major difference with this approach is that (1) the propagations from the concepts of the query are kept separate and that (2) they are not directly compared with the document. Rather, they are used to modify its semantic vector. In our experiments, our method gives better results. Also, we join [16] on their criticism of the propagation in a single vector, but our solutions are different.

Our approach also relies on the correspondences resulting from the matching of the two ontologies. Several existing matching algorithms could be used in our case [4]. In the interpretation step, we provide a very general algorithm to find the concept corresponding to the central concept of a SED. In case the concept is not shared, one could wonder whether matching algorithms could be used. In the solution we propose, the problem is quite different because the *weights of the concepts are also used* to find the corresponding concept (through the interpretation function). This is not the case in traditional ontology matching, which aim is to find general correspondences. In our case, one can see the problem as finding a "contextual" matching, the results of which cannot be used in other contexts. Because it is difficult to compute all the interpretation functions, one can use an *approximation algorithm* (for example, taking the least common ancestor as we did in our experiments). In that case, existing proposals can fit like [10, 14]. But it is clear that they do not find the best solution every time.

Finally, the word *interpretation* is used very often and reflects very different problems. However, to the best of our knowledge, it never refers to the case of interpreting a query expressed on some ontology, within the space of another ontology, by considering the weights of the concepts.

7. Conclusion

The main contribution of this paper is a proposal improving information exchange between a query initiator and a document provider that use different ontologies, in a context where semantic vectors are used to represent documents and queries. The approach only requires the initiator and the provider to share some concepts and also uses the unshared ones to find additional relevant documents. To our knowledge, the problem has never been addressed before and our approach is a first, encouraging solution. In short, 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. However, as far as shared concepts appear in a SED, expansion helps the document provider interpreting what the initiator wants, especially when the central concept is not shared. Interpretation by the document provider is not easy because the peers do not share the same vector space. Given its own ontology and similarity function, it first finds out a correspondent 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. Indeed if the effects of propagations from different central concepts were mixed in a single vector, the document provider wouldn't be able to interpret the query as precisely.

Although our approach builds on several notions (ontology, ontology matching, concept similarity, semantic indexing, relevance of a document wrt a query...) it is not stuck to a specific definition or implementation of them and seems compatible with many instantiations of them. It is important to notice that there is no human intervention at all in our experiments, in particular for semantic indexing. Clearly, in absolute, preci-

sion and recall could benefit from human interventions at different steps like indexation or the definition of the SEDs. Results show that our approach significantly improves the 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 test our approach in several different contexts in order to verify its robustness. Many different parameters can be changed: similarity and propagation functions, ontologies, indexing methods, corpus... Complexity is another point that should be considered carefully. Indeed, naive implementations would lead to unacceptable execution time. Although an implementation is running for the experiments within admissible times, it could benefit from a more thorough study of theoretical complexity.

References

[1] M. W. Berry, Z. Drmac, and E. R. Jessup. Matrices, vector spaces, and information retrieval. *SIAM Rev.*, 41(2), 1999.

[2] A. Bidault, C. Froidevaux, and B. Safar. Repairing queries in a mediator approach. In *ECAI*, 2000.

[3] E. Desmontils and C. Jacquin. *The Emerging Semantic Web*, chapter Indexing a web site with a terminology oriented ontology. 2002.

[4] J. Euzenat and P. Shvaiko. *Ontology matching*. Springer-Verlag, Heidelberg (DE), 2007.

[5] C. Fellbaum. *WordNet : an electronic lexical database*. 1998.

[6] T. Gaasterland, P. Godfrey, and J. Minker. An overview of cooperative answering. *J. of Intelligent Information Systems*, 1(2):123–157, 1992.

[7] A. Gómez-Pérez, M. Fernández, and O. Corcho. *Ontological Engineering*. Springer-Verlag, London, 2004.

[8] J. Gonzalo, F. Verdejo , I. Chugur, and J. Cigarran. Indexing with wordnet synsets can improve text retrieval. In *COLING/ACL '98 Workshop on Usage of WordNet for NLP*, 1998.

[9] 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.

[10] G. Jiang, G. Cybenko, V. Kashyap, and J. A. Hendler. Semantic interoperability and information fluidity. *Int. J. of cooperative Information Systems*, 15(1):1–21, 2006.

[11] R. Krovetz and W. B. Croft. Lexical ambiguity and information retrieval. *Information Systems*, 1992.

[12] D. Lin. An information-theoretic definition of similarity. In *International Conf. on Machine Learning*, 1998.

[13] R. Mandala and T. Tokunaga. Combining multiple evidence from different types of thesaurus for query expansion. In *SIGIR*, pages 191–197, 1999.

[14] E. Mena, A. Illaramendi, V. Kashyap, and A. Sheth. Observer: An approach for query processing in global information ssytems based on interoperation across preexisting ontologies. *Int. J. distributed and Parallel Databases*, 8(2):223–271, 2000.

[15] G. A. Miller and W. G. Charles. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 1991.

[16] 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.

[17] Y. Qiu and H. P. Frei. Concept based query expansion. In *SIGIR*, 1993.

[18] M.-C. Rousset. Small can be beautiful in the semantic web. In *International Semantic Web Conference*, pages 6–16, 2004.

[19] M. Sanderson. Retrieving with good sense. *Information Retrieval*, 2000.

[20] N. Seco, T. Veale, and J. Hayes. An intrinsic information content metric for semantic similarity in wordnet. In *ECAI*, 2004.

[21] C. Tempich, H. S. Pinto, and S. Staab. Ontology engineering revisited: An iterative case study. In *ESWC*, pages 110–124, 2006.

[22] A. Tversky. Features of similarity. *Psychological Review*, 84(4), 1977.

[23] E. M. Voorhees. Query expansion using lexical-semantic relations. In *SIGIR*, Dublin, 1994.

[24] W. Woods. Conceptual indexing: A better way to organize knowledge. Technical report, Sun Microsystems Laboratories, 1997.

[25] Z. Wu and M. Palmer. Verb semantics and lexical selection. In *ACL*, 1994.

[26] L. A. Zadeh. Fuzzy sets. *Information and Control*, 8, 1965.