

Perspectives on Ontology-based Querying

Rasmus Knappe, Henrik Bulskov, and Troels Andreassen

Department of Computer Science,
Roskilde University,
P.O. Box 260, DK-4000 Roskilde, Denmark
{knappe,bulskov,troels}@ruc.dk

Abstract. In this paper, we introduce principles for ontology-based querying of information bases. We consider a framework in which a basis ontology over atomic concepts in combination with a concept language defines a generative ontology. Concepts are assumed to be the basis for an index of the information base, in the sense that these concepts are attached to objects in the information base. Concepts are thus applied to obtain a means for descriptions that generalize classical word-based information base indexing. We discuss how the ontology influences the matching of values, especially how the different relations of the ontology may contribute to overall similarity between concepts. Further, we discuss a set of major properties to improve a given similarity measure's accordance with the semantics of the ontology, and use these properties to guide the choice of function.

Keywords: Fuzzy similarity, Flexible Querying, Ontology

1 Introduction

The approach presented here concerns ontology-based querying of information bases. The aim is to use knowledge from a domain-specific ontology covering the domain of a given information base to obtain better and closer answers on a semantic basis. Better answers are, in this context, primarily more fine-grained and better-ranked information base objects, which are obtained by exploiting better methods for computing the similarity between a query and objects from the information base.

We consider a generative ontology that defines a set of well-formed concepts from a basis ontology, which defines a vocabulary of concepts and situates these in a concept inclusion lattice. We assume an environment where queries as well as objects from the information base have well-formed concepts attached to them. This is an important aspect of the approach, because we can thereby reduce query evaluation to a matter of description comparison by bringing the descriptions of both objects and queries to a directly comparable form.

The environment for a system supporting this type of querying can be viewed therefore as a system that: 1) automatically produces conceptual descriptions,

used for conceptual indexing of text objects, and 2) supports natural language-based as well as word- list queries by initial transformation of these into descriptions for later comparison with the objects in the information base.

In this context, one of the major problems is to determine the similarity between the semantic elements.

It is no longer a simple match of keywords in the text objects, but also their meaning, that we must consider when we calculate the similarity between queries and objects in the information base.

2 A Generative Ontology

The purpose of the ontology is to define and relate concepts that can be used in descriptions of both queries and objects in the information base. The ontology framework is generative in the sense that a basis ontology defines a set of atomic concepts and situates these in a concept inclusion lattice, which is basically a taxonomy over single or multi-word concepts that are treated as atomic in the modeling of the domain. In combination with the given basis ontology, a concept language (description language) defines a set of well-formed concepts, used for description of both queries and objects in the information base.

The concept language in focus here, ONTOLOG[6], defines a set of semantic relations that can be used for “attribution” (feature-attachment) of concepts to form compound concepts. The number of available relations may vary with different domains, but among the more important relations that probably will be present in most domain models are the relations shown in Table 1

Table 1. Some important semantic relations used in most domain models.

Semantic relations
caused by (CBY)
characterized by (CHR)
concept inclusion (ISA)
with respect to (WRT)
temporal (TMP)
location (LOC)

Expressions in ONTOLOG are concepts situated in an ontology formed by an algebraic lattice with concept inclusion (ISA) as the ordering relation.

Attribution of concepts – combining atomic concepts into compound concepts by attaching attributes – can be written as feature structures. Simple attribution of a concept c_1 with relation r and a concept c_2 is denoted $c_1[r: c_2]$. For example, the concept *dog* can be given the characteristic (CHR) of having the color *black*, thereby forming the compound concept $dog[CHR: black]$, denoting the concept “*black dog*”.

We assume a set of atomic concepts \mathbf{A} and a set of semantic relations \mathbf{R} , as indicated with $\mathbf{R}=\{\text{WRT, CHR, CBY, TMP, LOC, \dots}\}$. Then the set of well-formed terms \mathbf{L} of the ONTOLOG language is recursively defined as follows.

- if $x \in \mathbf{A}$ then $x \in \mathbf{L}$
- if $x \in \mathbf{L}$, $r_i \in \mathbf{R}$ and $y_i \in \mathbf{L}, i = 1, \dots, n$
then $x[r_1: y_1, \dots, r_n: y_n] \in \mathbf{L}$

It appears that compound terms can be built from nesting, for instance, $c_1[r_1: c_2[r_2: c_3]]$ and from multiple attribution as in $c_1[r_1: c_2, r_2: c_3]$. For example, the sentence “*the dark gray cat*” can be interpreted as the nested semantic expression $cat[\text{CHR: gray}[\text{CHR: dark}]]$ or the multiple attributed expression $cat[\text{CHR: gray, CHR: dark}]$. The attributes of a term with multiple attributes $T = x[r_1: y_1, \dots, r_n: y_n]$ are considered as a set, thus we can rewrite T with any permutation of $\{r_1: y_1, \dots, r_n: y_n\}$.

The backbone of the ontology is the simple taxonomic concept inclusion relation ISA_{KB} , which is atomic in the sense that it defines a relation over the set of atomic concepts \mathbf{A} . It is considered as domain or world knowledge, and, for instance, may express the view of a domain expert or a knowledge engineer. We distinguish this (knowledge base) relation ISA_{KB} because concepts are assumed to be related by specific knowledge over the domain. For that reason we cannot expect the relation to be transitively closed. We define the relation ISA as the transitive closure of ISA_{KB} , while the relation $\text{ISA}_{\text{REDUC}}$ is the transitive reduction of ISA_{KB} .

Based on ISA , the transitive closure of ISA_{KB} , we can generalize into a relation over all well-formed terms of the language \mathbf{L} by the following:

- if $x \text{ ISA } y$ then $x \leq y$
- if $x[\dots] \leq y[\dots]$ then also
 $x[\dots, r: z] \leq y[\dots]$, and
 $x[\dots, r: z] \leq y[\dots, r: z]$,
- if $x \leq y$ then also
 $z[\dots, r: x] \leq z[\dots, r: y]$

where repeated \dots in each inequality denote zero or more attributes of the form $r_i: w_i$.

The basis ontology over atomic concepts, in combination with the formation rules for well-formed concepts in the concept language thereby forms the basis for a generative ontology. The concept language is introduced in order to be able to describe fragments of meaning in text more thoroughly than they can be described by simple keywords, while still refraining from a full- meaning representation, which is obviously not realistic in general search applications (with a huge database). A key question in the framework of querying is of course the definitions of similarity or nearness of terms, now that we no longer can rely on simple matching of keywords.

3 Ontology-based Similarity

In building a query evaluation principle that incorporates the knowledge represented in an ontology, a key issue is how the ontology influences the matching of values, that is, how the different relations of the ontology may contribute to similarity.

Therefore we must decide for each relation to what extent related values are similar and we must build similarity functions, mapping values into similarities, that reflect these decisions.

The following section motivates the choice of similarity functions by analyzing different principles for utilizing the structure of the ontology in devising similarity measures. Each principle is also related to previous work.

We describe firstly a shortest-path approach [4] to similarity based on the key ordering relation in the ontology, ISA_{KB} based on a definition for atomic concepts of the basis ontology we discuss how to extend the notion of similarity to cover general compound concepts as expressions in the language `ONTOLOG`.

Secondly, we introduce an alternative approach for devising a similarity measure based on the notion of shared nodes in a so-called similarity graph [5]. This approach can be considered as taking into account not only the shortest path but in principle all possible paths connecting two concepts.

Finally, we discuss a set of major properties that ensure that a given similarity function accords with the semantics of the ontology, and use these properties to guide the choice of function.

3.1 Shortest-path similarity on atomic concepts

The concept inclusion relation plays a central role as the ordering relation that binds the ontology in a lattice structure. Concept inclusion intuitively implies strong similarity in the opposite direction to inclusion (specialization). In addition, the direction of the inclusion (generalization) must, from a querying point of view, contribute with some degree of similarity. Take as an example the small fraction of an ontology in Figure 1. With reference to this ontology, the atomic concept *dog* can be directly expanded to cover also *poodle* and *alsatian*.

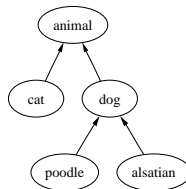


Fig. 1. Inclusion relation (ISA_{KB}) with upwards reading, e.g., *dog* ISA_{KB} *animal*.

This expansion respects the ontology in the sense that every concept subsumed by expanded concept *dog* by definition, every instance bears the relation ISA to *dog*. The intuition is that, to a query on *dog*, an answer including instances *poodle* is satisfactory (a specific answer to a general query). Because the hyponymy relation obviously is transitive we can by the same argument expand to further specializations, e.g., to include *poodle* in the extension of *animal*. However, similarity exploiting the lattice should also reflect “distance” in the relation. Intuitively greater distance (longer path in the relation graph) corresponds to smaller similarity with respect to the user’s query.

Further, generalization should contribute to similarity. Of course it is not strictly correct in an ontological sense to expand the concept *dog* with instances of *animal*, but because all *dogs* are *animals*, *animals* are to some degree similar to *dogs*. Thus, the property of generalization similarity should be exploited; but, for similar reasons as in the case of specializations, transitive generalizations should contribute a decreased degree of similarity.

A concept inclusion relation can be mapped into a similarity function in accordance with the intuition described above, as follows. Assume an ontology given as a domain knowledge relation ISA_{KB} . Figure 1 shows an example. The corresponding transitive closure relation ISA also includes, for instance, *poodle* ISA *animal*. To make “distance” influence similarity, we must consider the transitively reduced relation ISA_{REDUC} . Similarity reflecting distance can then be measured from path length in the graph corresponding to the ISA_{REDUC} relation of ISA_{KB} . A similarity function *sim* based on distance in ISA_{REDUC} $dist(X, Y)$ should have the properties:

1. *sim*: $U \times U \rightarrow [0, 1]$, where U is the universe of concepts
2. $sim(x, y) = 1$ only if $x = y$
3. $sim(x, y) < sim(x, z)$ if $dist(x, y) > dist(x, z)$

By parameterizing with two factors δ and γ , expressing similarity of immediate specialization and generalization, respectively, we can define a simple similarity function as follows. If there is a path between nodes (concepts) x and y in the hyponymy relation then it has the form

$$P = (p_1, \dots, p_n)$$

where

$$p_i \text{ } ISA_{REDUC} \text{ } p_{i+1} \text{ or } p_{i+1} \text{ } ISA_{REDUC} \text{ } p_i$$

for each i with $x = p_1$ and $y = p_n$.

Given a path $P = (p_1, \dots, p_n)$, set $s(P)$ and $g(P)$ to the numbers of specializations and generalizations, respectively, along the path P thus:

$$s(P) = |\{i | p_i \text{ } ISA_{REDUC} \text{ } p_{i+1}\}|$$

and

$$g(P) = |\{i|p_{i+1} \text{ ISA}_{\text{REDUC}} p_i\}|$$

If P^1, \dots, P^m are all paths connecting x and y , then the degree to which y is similar to x can be defined as

$$\text{sim}(x, y) = \max_{j=1, \dots, m} \left\{ \sigma^{s(P^j)} \gamma^{g(P^j)} \right\} \quad (1)$$

This similarity can be considered as derived from the ontology by transforming the ontology into a directional weighted graph, with σ as downwards and γ as upwards weights, and with similarity derived as the product of the weights on the paths. Figure 2 shows the graph corresponding to the ontology in Figure 1.

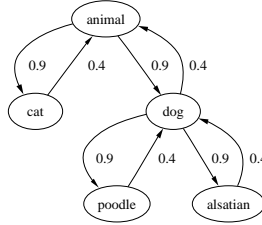


Fig. 2. The ontology transformed into a directed weighted graph, with the immediate specialization and generalization similarity values $\sigma = 0.9$ and $\gamma = 0.4$, respectively, as weights. Similarity is derived as the maximal (multiplicative) weighted path length, and thus $\text{sim}(\text{poodle}, \text{alsatian}) = 0.4 * 0.9 = 0.36$.

An atomic concept T can then be expanded to a fuzzy set, including T and similar values T_1, T_2, \dots, T_n as in:

$$T+ = 1/T + \text{sim}(T, T_1)/T_1 + \dots + \text{sim}(T, T_n)/T_n \quad (2)$$

Thus, for instance, with $\sigma = 0.9$ and $\gamma = 0.4$, the expansion of the concepts *dog*, *animal*, and *poodle* into sets of similar values would be:

$$\begin{aligned} \text{dog+} &= 1/\text{dog} + 0.9/\text{poodle} + 0.9/\text{alsatian} + 0.4/\text{animal} + 0.36/\text{cat} \\ \text{poodle+} &= 1/\text{poodle} + 0.4/\text{dog} + 0.36/\text{alsatian} + 0.16/\text{animal} + 0.144/\text{cat} \\ \text{animal+} &= 1/\text{animal} + 0.9/\text{cat} + 0.9/\text{dog} + 0.81/\text{poodle} + 0.81/\text{alsatian} \end{aligned}$$

The notion of deriving semantic relatedness or semantic similarity between concepts using network representations dates back to the spreading activation approach of Quillian [7] in the late 1960s.

Semantic similarity is classically viewed as a special case of semantic relatedness where we differentiate using only the defining features, typically equivalent to the ISA links between the concepts being compared [9]. Semantic relatedness between concepts can be viewed, on the other hand, as the aggregate of

the overall interconnection between the concepts in question, both defining and property-based.

The difference between relatedness and similarity can be illustrated by an example taken from Resnik, in which he argues that the relatedness between the concepts cars and gasoline seem to be higher than the relatedness between cars and bicycles, whereas the latter seem to be more similar [9].

The above approach to measuring similarity based on the shortest path between atomic concepts is therefore, using the Resnik’s distinction, a semantic similarity measure, rather than a measure of semantic relatedness. This measure shares similar aspects with the metric proposed by Rada et al. for use on a subset of semantic nets [8], where semantic similarity is derived using only the taxonomic (ISA) links of the semantic net.

The knowledge base in question for Rada et al.’s approach is formed by discriminating on the “criteriability” of the links – in other words, whether a link is between the concept and a defining feature or between the concept and a characteristic feature [8]. Rada et al. make the assumption that when only the ISA relation is used from the semantic net, then semantic relatedness and semantic similarity are equivalent. The motivation for their proposed semantic similarity measure, denoted *Distance*, is grounded in two observations. The first is that the behavior of conceptual distance should resemble that of a well-defined metric defined by a function $f(x, y)$ with the following properties [8]:

1. $f(x, x) = 0$, zero property,
2. $f(x, y) = f(y, x)$, symmetric property,
3. $f(x, y) \geq 0$, positive property, and
4. $f(x, y) + f(y, z) \geq f(x, z)$, triangular inequality

These criteria are equivalent to Tversky’s original axioms of minimality, which equals Rada et al.’s zero and positive property, the symmetry axiom, and the axiom of triangular inequality. The latter equals the symmetric property and the triangular inequality respectively [11].

The second assumption is that the conceptual distance between two concepts in a hierarchy is often proportional to the number of edges separating the concepts.

Rada et al.’s second assumption resembles the shortest-path similarity measure proposed, but the measure differs by being symmetric, which introduces inconvenient properties from a query expansion point-of-view, because a specific answer to a general question is better than a general answer to a specific question.

Resnik [9] introduces the notion of “information content” as the basis for an alternative way to measure semantic similarity in an ISA taxonomy. The measure combines the taxonomic structure with empirical probability estimates, to be capable of handling non-uniform distances in the ontology. Resnik argues that the intuitive similarity between concepts can be expressed by the extent to which they share information (defining features in the terms of Quillian), which is indicated by the presence of a concept that subsumes them both. This aspect

is partly covered in the approach by Rada et al., because the path between two very different but very specific concepts c_1 and c_2 will include very general concepts. This is because we will move high up in the ontology from c_1 before going down to c_2 – but this does not hold for the general case. Resnik defines the information content of a concept c , as the negative log likelihood, $-\log(p(c))$ of the probability of encountering c , and expresses the similarity between concepts c_1 and c_2 in terms of the maximal information content for the upper bounds (the set of concepts subsuming both c_1 and c_2):

$$sim(c_1, c_2) = \max_{c \in S(c_1, c_2)} [-\log(p(c))],$$

where $S(c_1, c_2)$ is the set of concepts that subsume both c_1 and c_2 .

This approach compensates for non-uniform distance in the ontology because the information content is non-decreasing as one moves up in the taxonomy, but it is symmetric, which is inconvenient from query expansion point-of-view, and it only considers the ordering ISA edges when computing similarity.

3.2 General shortest-path similarity

The semantic relations used to form concepts in the ontology indirectly contribute to similarity through subsumption. For instance, *noise*[CBY: *dog* [CHR: *black*]] is subsumed by each of the more general concepts *noise*[CBY: *dog*] and *noise*. Thus, with a definition of similarity covering atomic concepts, and in some sense reflecting the ordering relation (concept inclusion), we can extend to similarity on compound concepts by a relaxation that takes subsumed concepts into account when comparing descriptions.

The principle can be considered to be a matter of subsumption expansion. Any compound concept is expanded (or relaxed) into the set of subsuming concepts, thus:

noise[CBY: *dog*[CHR: *black*]]

is expanded to the set

$\{noise, noise[CBY: dog], noise[CBY: dog[CHR: black]]\}$

One approach to query-answering in this direction is to expand the description of the query using the ontology and the potential answer objects using subsumption.

For instance, a query on *dog* could be expanded to a query on similar values, such as:

$dog+ = 1/dog + \dots + 0.4/animal + \dots$

and a potential answer object such as *noise*[CBY: *dog*[WRT: *black*]] would then be expanded as exemplified above.

While not the key issue here, we should point out the importance of applying an appropriate averaging aggregation when comparing descriptions. For instance, the degree to which $c[r_1 : c_1]$ matches $c[r_1 : c_1[r_2 : c_2]]$ is higher than the degree to which c with no attributes matches $c[r_1 : c_1[r_2 : c_2]]$, and it is essential that similarity based on subsumption expansion exploits such relationships. Approaches to aggregation that can be tailored to obtain these properties, based on order-weighted averaging [12] and capturing nested structuring [13], are described in [1,2].

An alternative to the above subsumption expansion is to include edges that correspond to semantic relations in the computation of shortest-path similarity as a generalization of the principle of aggregating weights by multiplying cost factors, which was described in the previous subsection. This approach, considering the shortest path between concepts, which uses all semantic relations, is therefore a step in the direction of a similarity measure based on semantic relatedness. While the similarity between c and $c[r_1 : c_1]$ can be claimed to be justified by the ontology formalism (subsumption) or simply by the fact that $c[r_1 : c_1]$ ISA c , it is not strictly correct in an ontological sense to claim similarity likewise between c_1 and $c[r_1 : c_1]$.

For instance, $noise[CBY : dog]$ is conceptually not some kind of a *dog*. On the other hand, it would be reasonable to claim that $noise[CBY : dog]$ in a broad sense has something to do with (and thus has similarities to) *dog* (simply supported by the fact that concept $noise[CBY : dog]$ is present in the information base). Most examples tend to reveal the same characteristics, and this phenomenon is one good explanation for the comparative success of conventional word-based querying approaches. Basically, the (incorrect) assumption of no correlation between words in natural language phrases, which underlies any strictly word-based approach, does not lead to serious failure because the correlation that appears is not always dominating.

This could of course be an argument for not looking at compound concepts at all. Rather, these considerations point in the direction of re-assessing some of the importance of correlation in natural language phrases when developing similarity measures.

Consider Figure 3. The solid edges are ISA references and the broken ones are references by other semantic relations; in this example CBY and CHR are used. Each compound concept has broken edges to its attribution concept.

The principle of weighted path similarity can be generalized by introducing similarity factors for the semantic relations. The extensional arguments used to argue for differentiated weights depending on direction do not apply to semantic relations and seemingly there is no obvious way to differentiate based on direction at all. Thus one approach to the generalization is simply to introduce a single similarity factor and to transform to bidirectional edges.

Assume that we have k different semantic relations r^1, \dots, r^k , and let ρ_1, \dots, ρ_k be the attached similarity factors. Given a path $P = (p_1, \dots, p_n)$, set $r^j(P)$ to the number of r^j edges along the path P thus:

$$r^j(P) = |\{ i | p_i \overset{r^j}{\rightarrow} p_{i+1} \}| \quad (3)$$

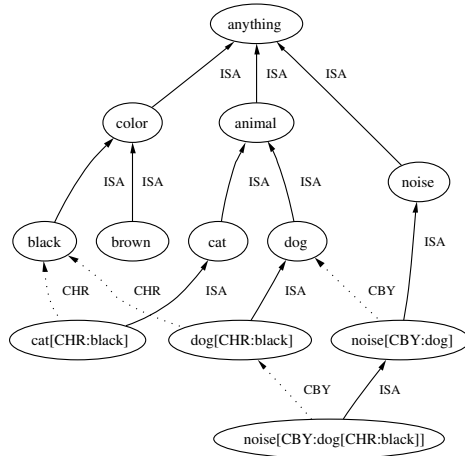


Fig. 3. An ontology where attribution with semantic relations is shown as dotted edges.

If P^1, \dots, P^m are all paths connecting c_1 and c_2 then the degree to which y is similar to x can be defined as:

$$\text{sim}(x, y) = \max_{j=1, \dots, m} \left\{ \sigma^{s(P^j)} \gamma^{g(P^j)} \rho_1^{r^1(P^j)} \dots \rho_k^{r^k(P^j)} \right\} \quad (4)$$

The result of transforming the ontology in Figure 3 is shown in 4. Here, two semantic relations CHR and CBY are in use. The corresponding edge count functions are r^{WRT} and r^{CBY} , and the attached similarity factors are denoted ρ_{WRT} and ρ_{CBY} . The figure shows the graph with the attached similarity factors as weights. Again, the degree to which a concept c_1 is similar to a concept c_2 is based on the shortest path (and is derived as the maximum of the products of edge weights over the set of paths connecting c_1 and c_2).

For instance, we can derive from Figure 4 that $\text{sim}(\text{cat}, \text{dog}) = 0.9 * 0.4 = 0.36$ and $\text{sim}(\text{cat}[\text{CHR}: \text{black}], \text{color}) = 0.2 * 0.4 = 0.08$.

The weights in the example are assigned in a rather ad hoc manner. Such assignment in practice would require careful consideration by domain experts. Furthermore, the similarity principle in general must be verified empirically.

Richardson et al. [10] present an approach to measuring similarity based on a combination of a weighted conceptual distance measure inspired by Rada et al. [8] and an information-based approach based on the work of Resnik [9] on the notion of information content. In this approach, Richardson et al. consider not only the key ordering relation ISA but also the meronym and holonym relations, for two reasons. Firstly, the part-whole relation contributes valuable information with respect to concept similarity. This is the argument for adapting an edge-counting method inspired by Rada et al., but considering all paths, weighted and aggregated in accordance with overall similarity. Secondly, the problems concerning non-uniform distances in the ontology and its sparsity should not be solved using only Resnik's measure of information content. We argue this on the

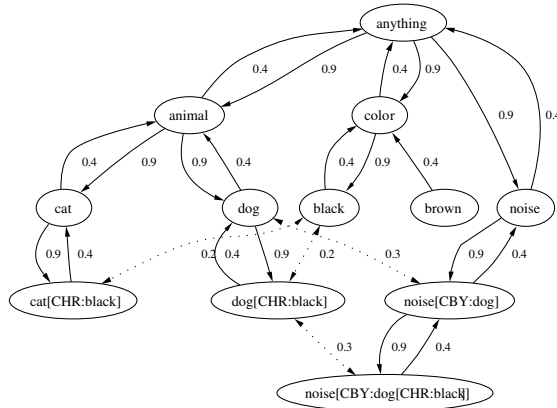


Fig. 4. The ontology of Figure 3 transformed into a directional weighted graph with the similarity factors for specialization: $\sigma = 0.9$, for generalization: $\gamma = 0.4$, for CBY: $\rho_{\text{CBY}} = 0.3$, and for CHR: $\rho_{\text{WRT}} = 0.2$.

basis of polysemous concepts. Take as an example the information content of the polysemous word “bank”, which will include all occurrences of bank regardless of meaning. This could be a problem, because it gives the same very high information content value to both “commercial bank” and “river bank”, which is intuitively wrong. The weighting of edges in Richardson et al.’s approach should depend on three criteria:

- the density of the graph at that point,
- the depth of the graph, and finally
- the strength of connotations between parent and child nodes

The authors thereby argue that a link in a dense area of the graph represents a smaller conceptual distance than a link in a less dense area. They argue further that increased depth implies increased similarity, which follows intuition because the similarity between *abstract* and *concrete* is smaller than the similarity between *cat* and *dog*. Finally, they consider the issue of non-uniform distance between parent and child at any level in the ontology. These criteria are a subset of the properties used to improve a given similarity measures accordance with the ontology, as described in the following section.

3.3 Shared nodes similarity

The shortest path approach described above is straightforward and does not entail computational problems. However, one aspect that must be assumed to contribute to similarity is ignored. When two concepts are connected by multiple paths only the shortest one contributes to similarity. Considering the ontology in Figure 4, the similarities between *cat[CHR:black]* and *dog[CHR:black]* according to the above method will not reflect the fact that they are both *black*.

This example shows that other connections than the shortest path should contribute in some cases, and it also indicates that similarity should be proportional to the number of possible paths connecting two concepts. Obviously, a similarity measure that takes into account all possible paths will impose increased computational complexity and calls for consideration of possible optimization approaches.

In this direction we suggest a similarity measure, based on the notion of a similarity graph [5], that utilizes a well-defined subset of all possible paths for measuring similarity.

A similarity graph can be viewed as a subpart of the ontology represented as a graph, with a subset of concepts as nodes and relations connecting these as edges. We define similarity graphs for any set of one or more concepts and specifically use the notion as a basis for similarity based on graph computations. The similarity between two concepts can thus be derived from a similarity graph covering these concepts.

In broader terms, our simplified “all-possible-paths” approach is a “shared nodes” approach, where shared nodes between two concepts are nodes that are “upwards reachable” from both concepts and where the similarity is dependent on the number of shared nodes.

To this end we define first the term decomposition $\tau(c)$ and the upwards expansion $\omega(c)$ of a concept term c . The term decomposition is defined as the set of all subterms of c , which thus includes all concepts subsuming c and all attributes of subsuming concepts for c . The term decomposition is defined as follows:

$$\tau(c) = \{x|c \leq x \vee c \leq y[r: x], x \in \mathbf{L}, y \in \mathbf{L}, r \in \mathbf{R}\}$$

As an example, the term *noise*[CBY: *dog*[CHR: *black*]] decomposes to:

$$\begin{aligned} \tau(\textit{noise}[\textit{CBY}: \textit{dog}[\textit{CHR}: \textit{black}]]) = \\ \{\textit{noise}[\textit{CBY}: \textit{dog}[\textit{CHR}: \textit{black}]], \textit{noise}[\textit{CBY}: \textit{dog}], \\ \textit{noise}, \textit{dog}[\textit{CHR}: \textit{black}], \textit{dog}, \textit{black}\} \end{aligned}$$

The upwards expansion $\omega(C)$ of a set of terms C is the transitive closure of C with respect to ISA_{KB} .

$$\omega(C) = \{x|x \in C \vee y \in C, y \text{ ISA } x\}$$

Thus, this expansion only adds atoms to C .

Now, a similarity graph $\gamma(C)$ for a set of concepts $C = \{c_1, \dots, c_n\}$ is defined as the graph that appears when decomposing C and connecting the resulting set of terms with edges corresponding to the ISA_{KB} relation and to the semantic relations used in attribution of elements in C . We define the triple (x, y, r) as the edge of type r from concept x to concept y .

$$\begin{aligned} \gamma(C) = \cup \\ \{(x, y, \text{ISA})|x, y \in \omega(\tau(C)), x \text{ ISA}_{\text{REDUC}} y\} \\ \{(x, y, r)|x, y \in \omega(\tau(C)), r \in \mathbf{R}, x[r: y] \in \tau(C)\} \end{aligned}$$

Figure 5 shows an example of such a similarity graph spanned by the two terms $poodle[CHR: black]$ and $cat[CHR: black]$.

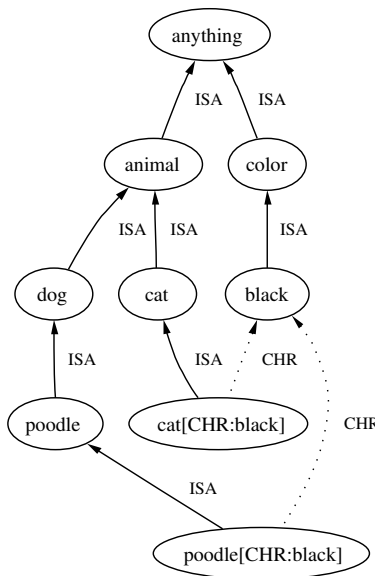


Fig. 5. The similarity graph for $poodle[CHR: black]$ and $cat[CHR: black]$.

One obvious approach for measuring similarity is to consider all possible connections between the concepts x and y . We may for instance have concepts connected directly through inclusion and in addition through an attribute dimension, as $cat[CHR: black]$ and $poodle[CHR: black]$. Taking all possible paths connecting two concepts x and y solves this problem, but involves a substantial increase in complexity. If we can reflect the multiple connections phenomenon without traversing all possible paths, we may have a more realistic means of similarity derivation. One option in this direction is to put emphasis on the nodes “shared” by x and y .

With $\alpha(x) = \omega(\tau(x))$ as the set of nodes (upwards) reachable from x in the similarity graph, we have $\alpha(x) \cap \alpha(y)$ as the reachable nodes shared by x and y , which thus obviously is an indication of what is common between x and y . Immediate transformations of this into a normalized similarity measure are the fractions of the cardinality of the intersection and the cardinality of, respectively, the union $\alpha(x) \cup \alpha(y)$ and the individual $\alpha(x)$ and $\alpha(y)$, giving the following normalized measures:

(a)

$$sim(x, y) = \frac{|\alpha(x) \cap \alpha(y)|}{|\alpha(x) \cup \alpha(y)|}$$

(b)

$$sim(x, y) = \frac{|\alpha(x) \cap \alpha(y)|}{|\alpha(x)|}$$

(c)

$$sim(x, y) = \frac{|\alpha(x) \cap \alpha(y)|}{|\alpha(y)|}$$

It is clear that similarity graphs and functions such as the above offer only a very coarse-grained approximation of whatever the genuine similarity may be. However the fact that the similarity is coarse is in itself typically an advantage rather than a problem in connection with querying, as long as the measure respects or “goes in the same direction as” the semantics. In the discussion of similarity functions below, we attempt to encircle major properties that ensure that a given function accords with the semantics of the ontology and use these properties to guide the choice of function.

First, it is important to notice that the similarity graph principle unifies the concept inclusion relation with the semantic relations used in attribution. We consider not only the ordering relation, but also take related concepts into consideration when calculating similarity. We therefore take not only *cat*, but also *black* to be related to *cat*[*CHR* : *black*] and we even take *cat*[*CHR* : *black*] to be related to *accident*[*CBY* : *cat*[*CHR* : *black*]].

A major property to guide the choice of similarity functions is:

Generalization cost property – the “cost” of generalization should be significantly higher than the cost of specialization.

The intuition is that, for instance, a *cat* satisfies the intention of an *animal* whereas an *animal* (that could be of any kind) does not necessarily satisfy the intention of a *cat*. From this property alone we can eliminate the first alternative similarity function (a) above. A consequence of insisting on this property is that the similarity function cannot be symmetrical, which (a) obviously is. In Figure 6 we should have $sim(D, B) < sim(B, D)$, according to the generalization cost property.

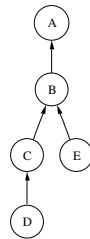


Fig. 6. *Generalization cost property* implies $sim(D, E) < sim(E, D)$ and *Specialization cost property* implies $sim(C, E) > sim(D, E)$

Now consider alternative (c). We see in Figure 6 that $sim(D, E) = \frac{2}{3}$ and $sim(E, D) = \frac{2}{4}$, which also violates the specialization cost property. Thus the only alternative that obeys the property is (b). With the example in Figure 6 we obtain $sim(D, E) = \frac{2}{4}$ and $sim(E, D) = \frac{2}{3}$

A second property that tends to appear more as optional because it is not implied by the semantics of the ontology is the following:

Specificity cost property – the “cost” of traversing edges should be lower when the connected nodes are more specific.

The intuition for this property is that the similarity between, for instance, siblings on low levels in the ontology such as “alsatian” and “poodle” should be higher than the similarity between siblings close to the top as “concrete” and “abstract”. This idea corresponds to the notion of information content described in [9], because the information content for *anything* (which subsumes both *physical* and *abstract*) is lower than the information content for *dog*.

Thus in Figure 7 we should have $sim(C, D) > sim(A, B)$. The similarity function (b) above satisfies this property: we have $sim(A, B) = \frac{2}{3}$ while $sim(C, D) = \frac{4}{5}$. ((a) and (c) also satisfy this property.)

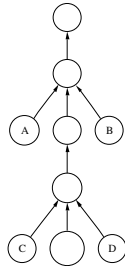


Fig. 7. *Specificity cost property*: implies that $sim(C, D) > sim(A, B)$.

A third property that similarly cannot be claimed to be semantically implied is the following.

Specialization cost property – further specialization implies reduced similarity.

As support of the intuition for this property, consider again Figure 6. The similarity function (b) obviously does not satisfy this property because we have $sim(E, C) = sim(E, D)$, and for K at any level of specialization below D we still have $sim(E, D) = sim(E, K)$.

This motivates us to consider alternative similarity functions that are influenced by both specialization and generalization (as the function (a) above is), but still do not violate the only ultimate property above, i.e., the generalization cost property (the anti-symmetry requirement). One modification that satisfies this is to simply take a weighted average of (b) and (c) above, as follows:

(d)

$$sim(x, y) = \rho \frac{|\alpha(x) \cap \alpha(y)|}{|\alpha(x)|} + (1 - \rho) \frac{|\alpha(x) \cap \alpha(y)|}{|\alpha(y)|}$$

where $\rho \in [0, 1]$ determines the degree of influence of generalizations.

Although simplicity favors similarity (b), and from the aspects discussed, this measure cannot be claimed to violate the semantics of the ontology, similarity (d) still appears to be a better choice. Similarity (b) is just a special case of (d) with $\rho = 1$. The parameter ρ allows us to tailor the similarity function, and can thereby comply with the generalization property.

As illustration consider the similarity graph in fig. 5. The similarities for *poodle*[*CHR : black*] and the other concepts included in the similarity graph are, when collected in a fuzzy subset of similar concepts (with $similar(x) = \Sigma sim(x, y)/y$) and $\rho = \frac{4}{5}$ the following:

$$\begin{aligned} similar(poodle[CHR : black]) = \\ 1.00/poodle[CHR : black] + 0,66/poodle + 0,59/cat[CHR : black] + \\ 0,54/dog + 0,54/black + 0,43/animal + 0,43/color + 0,36/cat + 0.31/anything \end{aligned}$$

The purpose of similarity measures in connection with querying is of course to look for similar rather than exactly matching values, that is, to introduce soft rather than crisp evaluation. As indicated through examples, one way to introduce similar values is to expand crisp values into fuzzy sets that include similar values. Expansion of this kind, applying similarity based on knowledge in the knowledge base, is a simplification replacing direct reasoning over the knowledge base during query evaluation. Graded similarity is the obvious means to make expansion useful, because by using a simple threshold values for similarity, the size of the answer can be fully controlled.

4 Conclusion

We have described principles for measuring similarity between atomic as well as compound concepts that draw on the structure and the relations of the ontology.

The notion of measuring similarity as distance, either in the ordering relation or in combination with the semantic relations, seems to indicate a usable theoretical foundation for design of similarity measures.

Similarity based on shared nodes in a similarity graph introduces a more fine-grained similarity between concepts, by considering shared attribution, but without adding significantly to the overall computational complexity.

Furthermore, when choosing a similarity function, it is important to ensure that it accords with the semantics of the ontology. Qualitative assessment with respect to a set of major properties appears to be a useful guide to the choice.

The expansion of crisp values into fuzzy sets that include similar values, is useful, because the size of the answer can be fully controlled by using simple threshold values for similarity.

Acknowledgments

The work described in this paper is part of the OntoQuery¹[3] project supported by the Danish Technical Research Council and the Danish IT University.

References

1. Andreasen, T.: On knowledge-guided fuzzy aggregation, 9th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, IPMU'2002, Annecy, France, July 1–5, 2002, Proceedings
2. Andreasen, T.: Query evaluation based on domain-specific ontologies. 20th IFSA / NAFIPS International Conference Fuzziness and Soft Computing, NAFIPS'2001, pp. 1844–1849, Vancouver, Canada, 2001, Proceedings
3. Andreasen, T.; Jensen, P. Anker; Nilsson, J. Fischer; Paggio, P.; Pedersen, B. Sandford & Thomsen, H. Erdman: Ontological Extraction of Content for Text Querying, NLDB'2002, Stockholm, Sweden, 2002.
4. Bulskov, H.; Knappe, R. & Andreasen, T.: On Measuring Similarity for Conceptual Querying, LNAI 2522, pp. 100–111 in Andreasen T.; Motro A.; Christiansen H.; Larsen H.L.(Eds.): Flexible Query Answering Systems 5th International Conference, FQAS 2002, Copenhagen, Denmark, October 27–29, 2002, Proceedings
5. Knappe, R.; Bulskov, H. & Andreasen, T.: Similarity Graphs, LNAI 2871, pp. 668–672 in Zhong N.; Ras Z.W.; Tsumoto S.; Suzuki E. (Eds.): 14th International Symposium on Methodologies for Intelligent Systems, ISMIS 2003, Maebashi, Japan, October 28–31, 2003, Proceedings
6. Nilsson, J. Fischer: A Logico-algebraic Framework for Ontologies – ONTOLOG, in Jensen, P. Anker & Skadhauge, P. (eds.): Proceedings of the First International OntoQuery Workshop – Ontology-based interpretation of NP's, Department of Business Communication and Information Science, University of Southern Denmark, Kolding, 2001
7. Quillian, M. R.: Semantic Memory, in Semantic Information Processing (Minsky, M., ed.), MIT Press, 1968
8. Rada, Roy; Mili, Hafeedh; Bicknell, Ellen & Blettner, Maria: Development and Application of a Metric on Semantic Nets, IEEE Transactions on Systems, Man, and Cybernetics, Volume 19, Number 1, pp. 17–30, 1989
9. Resnik, Philip: Semantic Similarity in a Taxonomy: An Information-based Measure and its Application to Problems of Ambiguity in Natural Language, Journal of Artificial Intelligence, pp. 95–130, 1999
10. Richardson, Ray; Smeaton, Alan F. & Murphy, John: Using WordNet as Knowledge Base for Measuring Semantic Similarity between Words. Technical Report CA-1294, Dublin City University, School of Computer Applications, 1994.
11. Tversky, Amos: Features of Similarity, Psychological Review, Volume 84, Number 4, pp. 327–352, 1977
12. Yager, R.R.: On ordered weighted averaging aggregation operators in multicriteria decision making, in IEEE Transactions on Systems, Man and Cybernetics, vol 18, 1988.

¹ The project has the following participating institutions: Centre for Language Technology, The Technical University of Denmark, Copenhagen Business School, and Roskilde University.

13. Yager, R.R.: A hierarchical document retrieval language, in *Information Retrieval* vol 3, Issue 4, Kluwer Academic Publishers pp. 357–377, 2000.