

# On Similarity Measures for Concept-based Querying

Rasmus Knappe and 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

The aim of this paper is to devise measures for similarity for querying an information base, that utilizes the knowledge represented in an ontology. The basis for this approach is an ontology defining the major concepts of the domain and a concept language defining a set of well-formed concepts. These concepts are to be used for description of the semantics of objects in the information base. Well-formed concepts are thus assumed to form the basis for an indexing of the information base in the sense that these concepts appear as descriptions attached to the objects in the information base. The challenge for support of queries in this context is an evaluation principle that on the one hand utilizes the semantics expressed in the concept language and on the other is sufficiently efficient to candidate as a realistic principle for query evaluation. We present and discuss principles where efficiency is obtained by reducing the matching problem - which basically is a matter of conceptual reasoning - to numerical similarity computation.

## 1 Introduction

The objective of this paper is to devise similarity measures that utilizes knowledge from a domain-specific ontology to obtain better answers on a semantical level, thus comparing concepts rather than words. The basis is an ontology that defines and relates concepts and a concept language for expressing the semantics

of queries and objects in the information base. We consider an environment where queries and objects in the information base are attached with descriptions (well-formed concepts), hence making query evaluation a matter of comparison of descriptions.

The approach presented in the paper is a refinement of our earlier work[3] on similarity measures based on distance in an ontology. We aim to devise a similarity measure that can capture an aspect not included in[3]. The intention is also to cover the intuition that for example the similarity between concepts “*grey cat*” and “*grey dog*” is intuitively higher than the similarity between “*grey cat*” and “*yellow bird*”, because the former share the same color.

We describe below a concept language that can express the needed semantics of concepts and thereby form the basis for a similarity measure that is able to take into account that concepts share attributes, as grey in the example above.

## 2 Concepts and concept language

The role of the ontology is to define and relate a set of atomic concepts by situating these in a concept inclusion lattice, which basically is a taxonomy over concepts. In combination with the ontology we have a concept language ONTOLOG[5] which is a concept algebra for integration, formalization, representation and reasoning with semantics of natural language and ontologies. Expressions in ONTOLOG are descriptions of concepts situated in the ontology. The latter is an important aspect of the concept language since it allows us to use the descriptions of concepts in combination with the ontology to measure similarity between concepts, hence between queries and objects in the information base. Hence an obvious way to

perform query evaluation is by means of description comparison.

The concept language ONTOLOG is based on a set of atomic concepts defined by the ontology. The language can combine concepts into compound concepts using semantic relations. The suitable collection of available semantic relations may vary with different domains. Among the more general relations are WRT (With-respect-to), CHR (Characterized-by), CBY (Caused-by), TMP (Temporal) and LOC (Location). Attribution of concepts, i.e. the combining of atomic concepts with semantic relations into compound concepts, can be written as feature structures. Simple attribution of a concept  $c_1$  with relation  $r$  and a concept  $c_2$  is thus denoted  $c_1[r: c_2]$ . An example is the ONTOLOG expression  $cat[CHR: grey]$  describing the concept “grey cat”.

We assume a set of atomic concepts  $\mathbf{A}$  and a set of semantic relations  $\mathbf{R}=\{WRT,CHR,\dots\}$ . 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  $A[r_1: B[r_2: C]]$  and from multiple attribution as in  $A[r_1: B, r_2: C]$ . Take as an example the sentence: “the dark grey cat” which can be interpreted as the nested semantic expression  $cat[CHR: grey[CHR: dark]]$  or the multiple attributed expression  $cat[CHR: grey, CHR: dark]$ . The attributes of a multiple attributed term  $T = x[r_1: y_1, \dots, r_n: y_n]$  is considered as a set, thus we can rewrite T with any permutation of  $\{r_1: y_1, \dots, r_n: y_n\}$ .

The basis for the ontology is a simple taxonomic concept inclusion relation  $ISA_{KB}$  that defines the hyponymy lattice over the set of atomic concepts  $\mathbf{A}$ . This relation is considered as domain or world knowledge and may for instance express the view of a domain expert.

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  $x ISA y$  then  $x \leq y$

- if  $x[r_1: y_1, \dots, r_n: y_n] \leq y$  then  $x[r_1: y_1, \dots, r_n: y_n, r_{n+1}: y_{n+1}] \leq y$

The purpose of the language introduced above is to describe fragments of meaning in text at a more thoroughly way than what can be obtained from simple keywords, while still refraining from full meaning representations which is obviously not realistic in general search applications (with a huge database).

### 3 Similarity

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.

As shown in [3] the shortest path from the ISA relation can be used as a measure for similarity. This approach is to transitively reduce the ISA relation (forming the relation denoted  $ISA_{REDUC}$ ) to make distance influence the computation of similarity.

Consider Figure 1. The solid edges are ISA references and the broken are references by other semantic relations - in this example only CHR. Each compound concept has broken edges to its attributed concepts.

Figure 1: A subontology covering colored pets

If we consider only the ISA-edges then there is no difference in similarity between any pair of  $dog[CHR: grey]$ ,  $cat[CHR: grey]$  and,  $bird[CHR: yellow]$  due to the fact that they are all specializations (sub-classes) of pet.

If we on the other hand also consider broken edges, then we can add the aspect of shared attribution to the computation of similarity and thus refine the measure. In the case of Figure 1 we can, by including the bro-

ken edges, capture the intuitive difference in similarity between two grey pets compared to the similarity between a grey and a yellow pet. This difference is visualized by existence of a path, that includes the shared concept, between the two concepts sharing attribution.

The general idea, in this paper, is therefore a refinement by defining a similarity measure between concepts  $c_1$  and  $c_2$  upon the set of all possible paths in the graph, representing the part of the ontology covering  $c_1$  and  $c_2$ .

The inclusion of attribution in the approach increases the overall complexity and we therefore devise a similarity measure that utilizes a well-defined subset of all possible paths, by using the notion of shared nodes between the compared concepts. The goal now is to encircle a basis in the form of a subontology for measuring similarity, thus reducing complexity.

To this end we define first the term-decomposition  $\tau(c)$  and the upwards expansion  $\varpi(c)$  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:

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

Consider as an example the decomposition of the term

$$\tau(\text{cat}[\text{CHR} : \text{grey}[\text{CHR} : \text{dark}]])$$

resulting in the set containing the following concepts:

$$\{ \text{cat}[\text{CHR} : \text{grey}[\text{CHR} : \text{dark}]], \\ \text{cat}[\text{CHR} : \text{grey}], \\ \text{cat}, \\ \text{grey}[\text{CHR} : \text{dark}], \\ \text{grey}, \\ \text{dark} \}$$

The upwards expansion  $\varpi(C)$  of a set of terms  $C$  is the transitive closure of  $C$  with respect to  $\text{ISA}_{\text{KB}}$ :

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

This expansion thus only adds atoms to  $C$ .

We define further the upwards spanning subgraph (subontology)  $\gamma(C)$  for a set of concepts  $C = \{c_1, \dots, c_n\}$  as the graph that appears when decomposing  $C$  and connecting the resulting set of terms

with edges corresponding to the  $\text{ISA}_{\text{KB}}$  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  $y$ .

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

Now a shared node between concepts  $c_1$  and  $c_2$  is a node that is reachable from both  $c_1$  and  $c_2$ . If we for a concept  $c$  define  $\alpha(c)$  to be the nodes reachable from  $c$ , that is  $\alpha(c) = \omega(\tau(c))$ , then  $\alpha(c_1) \cap \alpha(c_2)$  is the set of shared nodes for two concepts  $c_1$  and  $c_2$ .

With the example in Figure 1 both  $\{Grey, Color, Animal, \dots, Anything\}$  are shared nodes for concepts  $dog[\text{CHR} : grey]$  and  $cat[\text{CHR} : grey]$ . Whereas concepts  $dog[\text{CHR} : grey]$  and  $bird[\text{CHR} : yellow]$  does not share attribution and therefore share one node less, namely  $\{Grey\}$ .

As seen in Figure 1 the notion of shared nodes can be used to include attribution in the calculation of overall similarity between concepts.

The transformation of a similarity measure based on aggregation over possible paths between concepts into a measure based on the notion of shared nodes has, as a motivating factor the possible reduction of overall computational complexity.

## 4 Conclusion

We have described a principles for measuring similarity between both atomic and compound concepts that draws on meta knowledge.

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. Furthermore the notion of similarity based on shared nodes seems to nuance of overall similarity between concepts sharing attribution, without adding significantly to the overall computational complexity.

The purpose of similarity measures in connection with querying is of course to look for similar rather than for exactly matching values, that is, to introduce soft rather than crisp evaluation. As indicated through examples above one approach to introduce similar val-

ues is to expand crisp values into fuzzy sets including also similar values. Applying similarity based on knowledge in the knowledge base, is a simplification replacing direct reasoning over the knowledge base during query evaluation. The graded similarity is the obvious means to make expansion a useful - by using simple threshold values for similarity the size of the answer can be fully controlled.

## Acknowledgments

The work described in this paper is part of the OntoQuery[1, 2]<sup>1</sup> project supported by the Danish Technical Research Council.

## References

- [1] Andreasen, T., Nilsson, J. Fischer & Thomsen, H. Erdman: Ontology-based Querying, in Larsen, H.L. *et al.* (eds.) *Flexible Query Answering Systems, Flexible Query Answering Systems, Recent Advances*, Physica-Verlag, Springer, 2000. pp. 15-26.
- [2] Andreasen, T., Jensen, P. Anker, Nilsson, J. Fischer, Paggio, P, Pedersen, B. Sandford & Hanne Erdman Thomsen: Ontological Extraction of Content for Text Querying. In: Anderson, Birger; Maria Bergholtz & Paul Johanneson (eds.): *NLDB 2002, 7th International Workshop on Applications of Natural Language to Information Systems*, June 27-28, Stockholm. Printed in preproceedings, forthcoming in *Lecture Notes in Computer Science*, Springer-Verlag, 2002.
- [3] Bulskov, H., Knappe, R. and Andreasen, T.: On Measuring Similarity for Conceptual Querying, *LNAI 2522*, pp. 100-111 in T. Andreasen, A. Motro, H. Christiansen, H.L. Larsen (Eds.): *Flexible Query Answering Systems 5th International Conference, FQAS 2002*. Copenhagen, Denmark, October 27-29, 2002. Proceedings
- [4] Jensen, P. Anker & Skadhauge, P. (eds.): *Proceedings of the First International Onto-Query Workshop Ontology-based interpretation*

*of NP's*, Department of Business Communication and Information Science, University of Southern Denmark, Kolding, 2001, to be republished at [www.ontoquery.dk](http://www.ontoquery.dk).

- [5] Nilsson, J. Fischer: A Logico-algebraic Framework for Ontologies ONTOLOG, in [4].

---

<sup>1</sup>The project has the following participating institutions: Center for Language Technology, The Technical University of Denmark, Copenhagen Business School, Roskilde University, and the University of Southern Denmark.