

Similarity Graphs

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. The focus of this paper is approaches to measuring similarity for application in content-based query evaluation. Rather than only comparing at the level of words, the issue here is conceptual resemblance. The basis is a knowledge base defining major concepts of the domain and may include taxonomic and ontological domain knowledge. The challenge for support of queries in this context is an evaluation principle that on the one hand respects the formation rules for concepts 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 a similarity measure that utilizes knowledge from a domain-specific ontology to obtain better answers on a semantical level, thus comparing concepts rather than terms. Better answers are primarily better ranked information base objects which in turn is a matter of better means for computing the similarity between a query and an object from the information base. The basis is an ontology that defines and relates concepts and a concept language ONTOLOG [6] for expressing the semantics of queries and objects in the information base.

The approach presented in the paper is a refinement of [5] on similarity measures based on the notion of shared nodes. We aim to devise a similarity measure that can capture the aspect exemplified by 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.

This is sought done by introducing the notion of similarity graphs, as the subset of the ontology covering the concepts being compared, thereby capturing semantics without the loss of scalability.

2 Similarity Graphs

The basis for the ontology is a simple taxonomic concept inclusion relation ISA_{KB} which is considered as domain or world knowledge and may for instance express the view of a domain expert. The concepts in the ontology can be divided into two sets; atomic and compound concepts. The latter are formed by attribution of atomic concepts with a relation to form a compound concept. Take as an example the compound concept “black cat”, denoted $cat[CHR: black]$ in the concept language ONTOLOG.

The general idea now is a similarity measure between concepts c_1 and c_2 based upon the set of all nodes reachable from both concepts in the similarity graph, representing the part of the ontology covering c_1 and c_2 . These shared nodes reflect the similarity between concepts, both in terms of subsuming concepts and similar attribution.

Consider Figure 1. The solid edges are ISA_{KB} 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.

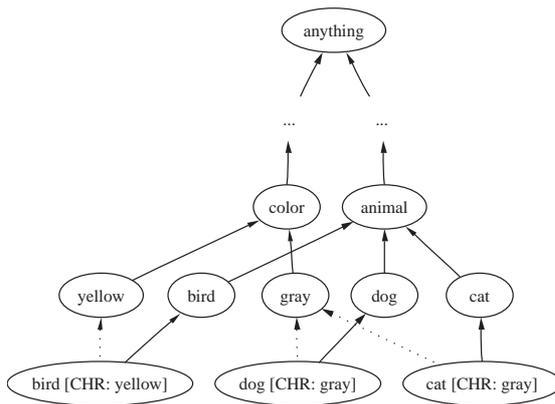


Fig. 1. An ontology covering colored pets.

If we consider only the ISA_{KB} -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. In the case of Figure 1 we can, by including the broken 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 the existence of a path, that includes the shared concept, between the two concepts sharing attribution.

One possible way to include the attributes of compound concepts in the basis for a similarity graph is to perform a term-decomposition of the concepts. To further include the semantics, given by the inclusion relation of the ontology we can expand every atomic concept in the decomposition with all nodes upwards reachable in the ontology.

The term-decomposition is defined as the set of all terms appearing in c . The present term-decomposition differs from earlier definitions [5] in that it does not generate all generative subsuming concepts. We only need to maintain the compound nestings and the atomic concepts of a given concept when decomposing. For instance, we do not need to include $noise[CBY: dog]$ in the decomposition of $noise[CBY: dog[CHR: black]]$, because this concept will be included in any specialization of $noise[CBY: dog]$.

If we for a concept $c = c_0[r_1 : c_1, \dots, r_n : c_n]$, where c_0 is the atomic concept attributed in c and c_1, \dots, c_n are the attributes (which are atomic concepts or further compound concepts), define:

$$subterm(c) = \{c_0, c_1, \dots, c_n\}$$

and straightforwardly extend $subterm$ to be defined on a set of concepts $C = \{c_1, \dots, c_n\}$, such that

$$subterm(C) = \cup_i subterm(c_i)$$

then we can obtain the term-decomposition of c as the closure by subterm, that is, by repeatedly applying $subterm$:

$$\tau(c) = \{c\} \cup \{x | x \in subterm^k(c) \text{ for some } k\}$$

As an example the concept $noise[CBY: dog[CHR: black]]$ decomposes to the following set of concepts:

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

The upwards expansion, i.e. nodes upwards reachable in the ontology, $\omega(C)$ of a set of concepts C is then the transitive closure of C with respect to ISA_{KB} .

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

where ISA is the transitive closure of ISA_{KB} . This expansion thus only adds atoms to C .

Now a similarity graph $\gamma(C)$ is defined 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 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 .

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

Fig. 2 shows an example of a similarity graph covering two terms $cat[CHR: black]$ and $poodle[CHR: black]$, capturing similar attributes and subsuming concepts.

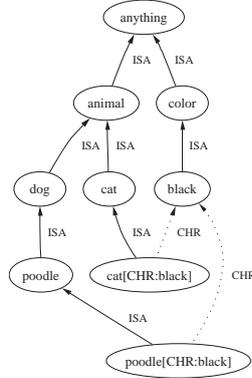


Fig. 2. An example of a similarity graph for the concepts $cat[chr: black]$ and $poodle[chr: black]$.

A similarity graph expresses the similarity between two concepts c_1 and c_2 . We can therefore, by generalizing to every pair of concepts in an ontology, identify similar concepts $\{c_1, \dots, c_n\}$ to a given concept c . This set can be expressed as a fuzzy set where the membership grade for a given element w_i/c_i is the similarity between c and c_i .

3 Conclusion

The notion of measuring similarity based on the notion of similarity graphs, seems to indicate a usable theoretical foundation for design of similarity measures. The inclusion of the attribution of concepts, by means of shared nodes, in the calculation of similarity, gives a possible approach for a measure that captures more details and at the same time scale to large systems.

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 values is to expand crisp values into fuzzy sets including also similar values. Query evaluation in such an environment raises the need for aggregation principles that supports nested aggregation over fuzzy sets [7] as described in [1].

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. 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 OntoQuery¹[4] project supported by the Danish Technical Research Council and the Danish IT University.

References

- [1] Andreasen, T.: On knowledge-guided fuzzy aggregation. In *IPMU'2002, 9th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, 1-5 July 2002, Annecy, France
- [2] Andreasen, T.: Query evaluation based on domain-specific ontologies. In *NAFIPS'2001, 20th IFSA / NAFIPS International Conference Fuzziness and Soft Computing*, pp. 1844-1849, Vancouver, Canada, 2001.
- [3] 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.
- [4] Andreasen, T., Jensen, P. Anker, Nilsson, J. Fischer, Paggio, P., Pedersen, B. Sandford & Thomsen, H. Erdman: Ontological Extraction of Content for Text Querying, to appear in *NLDB 2002*, Stockholm, Sweden, 2002.
- [5] Knappe, R., Bulskov, H. and Andreasen, T.: On Similarity Measures for Content-based Querying, LNAI, to appear in *International Fuzzy Sysytems Association, World Congress*, June 29-July 2, Istanbul, Turkey, 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] Yager, R.R.: A hierarchical document retrieval language, in *Information Retrieval* vol 3, Issue 4, Kluwer Academic Publishers pp. 357-377, 2000.

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