

On Measuring Similarity for Conceptual Querying

Henrik Bulskov, Rasmus Knappe, and Troels Andreasen

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

Abstract. The focus of this paper is approaches to measuring similarity for application in connection with query evaluation. Rather than only comparing at the level of words the issue here is to compare concepts that appear as compound expressions derived from list of words through brief natural language analysis. Concepts refers to and are compared with respect to an ontology describing the domain of the database. We discuss three different principles for measuring similarity between concepts. One in the form of subsumption expansion of concepts and two as different measures of distance in a graphical representation of an ontology.

1 Introduction

The magnitude and complexity of available information sources constitutes a rapidly increasing conglomerate, that requires both a general overview and domain specific insight knowledge in order to query and utilize.

While classical term based text retrieval systems produce answers of a relatively high quality, considering the simplicity of the methods involved, there definitely appears to be room for improvement. Especially when dealing with the motivating problem in classical term based text retrieval systems – their obvious inability to interpret queries as anything other than a list of terms. This constitutes a problem because the semantics of the text fragment is lost along with the context and the semantic relations between concepts [7].

We are looking for improvement that can enhance the systems ability to generate ideal answers.

1.1 The OntoQuery Project

The foundation for the work in this paper is the interdisciplinary research project ONTOQUERY¹ (Ontology-based Querying)[3, 4, 10]. The primary goal for the

¹ The project has the following participating institutions: Centre for Language Technology, The Technical University of Denmark - Informatics and Mathematical Modelling, Copenhagen Business School - Data linguistics, Roskilde University - Intelligent Systems Laboratory and the University of Southern Denmark.

ONTOQUERY project is to contribute to the development of theories and methodologies for content-based text retrieval from text databases.

This is sought done by introducing:

- A formal description language, ONTOLOG [9], whose expressions functions as descriptions for concepts in the domain texts, in the queries and in the ontology.
- A method for doing ontology-based linguistic analysis.
- Theories for ontology-based query processing that efficient compares descriptions of queries with descriptions of elements in the text domain [2, 1].

Along with the work on theoretical issues a prototype system with a set of accompanying tools is developed for validation and demonstration of the theoretical results [5].

2 Object Representation in the OntoQuery project

An important aspect in developing information retrieval systems concerns the comparing of queries with the objects held by the system.

One approach is a common representation, for both queries and objects, in which we bring the description of the query and the description of the objects on a directly comparable form.

A central aspect of this approach is that descriptions are created as intermediate representations of “*content*”. A description is a set of descriptors describing a text fragment. For a text fragment, e.g. a sentence, a simple form description expresses the content by means of a nested set of words from the sentence.

Descriptions have the general form:

$$D = \{D_1, \dots, D_n\} = \{\{D_{11}, \dots, D_{1m}\}, \dots, \{D_{n1}, D_{n2}, \dots, D_{nm}\}\}$$

where each descriptor D_i is a set of concepts D_{i1}, \dots, D_{im} . This representation is shown in figure 1.

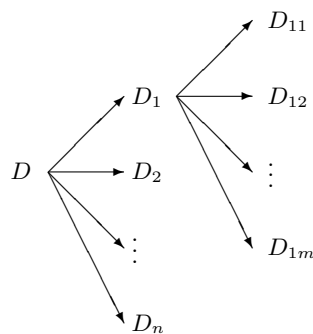


Fig. 1. The nested set representation.

To generate descriptions, text fragments are prepared by a parser that employ the knowledge base. The parser can in principle scale from a simple word recognizer to a complex natural language parser that maps the full meaning content of the sentence into an internal representation.

Since the issue here is IR the idea is of course to grab fragments of content rather than represent full meaning, and the building stones are concepts. The structure of descriptions should be understood considering this aim.

The approach to description generation in the ONTOQUERY project is a subject to ongoing development. In the present state descriptions are generated as follows.

A tagger identifies heuristically categories for words. Based on tags and a simple grammar, a parser divides the sentence by framing identified noun phrases (NPs) in the sentence. For each part of the sentence, corresponding to an NP, a descriptor is produced as a set of concepts.

Take as an example the sentence:

“Physical well-being caused by a balanced diet”

A description, consisting of nouns and adjectives, in a simple form without taking into account the framing of NP’s in the sentence could be:

(physical), (well-being), (balanced), (diet)

With the framing of NP’s the descriptors can be gathered giving the description:

(physical well-being), (balanced diet)

2.1 Introducing Semantic Relations

The representation described until now, and implemented in the present prototype in the ONTOQUERY project, is nested sets of concepts, as shown in Fig. 1. But not all the information from the linguistic analysis is mapped into this representation. For instance when representing “*balanced, diet*” as the set of two words {*balanced, diet*} the information about the semantic relation (obvious *characterized by*) is not preserved.

The experiments with the present prototype has shown that the nested set representation is suitable for describing the complexity of sentences and is efficient as model for the query evaluations.

Changing the innermost level in the nested sets from concept to semantic expressions, will preserve the properties of the nested set representation and at the same time increase the expressive power.

In the sentence “*Physical well-being caused by a balanced diet*” the nouns (e.g. concepts) are *well-being* and *diet* and the words indicating the semantic relation between these concepts are *caused by*. Without the semantic relation this sentence describes something about *well-being* and *diet*, but not exactly how they are connected. The semantic relations combine the concepts, constituting a

more complex concept. In this example the concept of physical well-being caused by a balanced diet.

The semantic relation in the above example is *caused by*. Natural language has an arbitrary set of semantic relations but in the IR context only a minor subset is relevant [8]. In examples in this paper we use the subset of semantic relations shown in Table 1

Table 1. The subset of semantic relations and their abbreviations used in this paper.

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

2.2 Semantic Expressions

A representation with nested sets of semantic expression needs a formal language to describe compound concepts formed by semantic relations.

ONTOLOG 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 an ontology formed by an algebraic lattice with concept inclusion as the ordering relation.

The basic elements in ONTOLOG are concepts and binary relations between concepts. The algebra introduces two closed operations on concept expressions φ and ψ [9]:

- conceptual *sum* ($\varphi + \psi$), interpreted as the concept being φ or ψ
- conceptual *product* ($\varphi \times \psi$), interpreted as the concept being φ and ψ

Relationships r are introduced algebraically by means of a binary operator ($:$), the Peirce product ($r : \varphi$), which combines a relation r with an expression φ , resulting in a expression, thus relating nodes in the ontology.

The typical use of the Peirce products is as a factor in conceptual products, as in $c \times (r : c_1)$, which can be rewritten to form the feature structure $c[r : c_1]$, where $[r : c_1]$ is an attribution of the concept c .

As an example consider the sentence "disorders caused by hormones". One possible resulting description could be:

$$disorder[CBY : hormones]$$

Nesting is supported as in:

$$disorder[CBY : lack[WRT : vitaminD]]$$

Describing a "disorder caused by lack of vitamin D". A concept c with multiple attributions is denoted:

$$c \begin{bmatrix} r_1 : \varphi_1 \\ \dots \\ r_n : \varphi_n \end{bmatrix}$$

Which translates into the ONTOLOG expression $c \times r_1(\varphi_1) \dots \times r_n(\varphi_n)$.

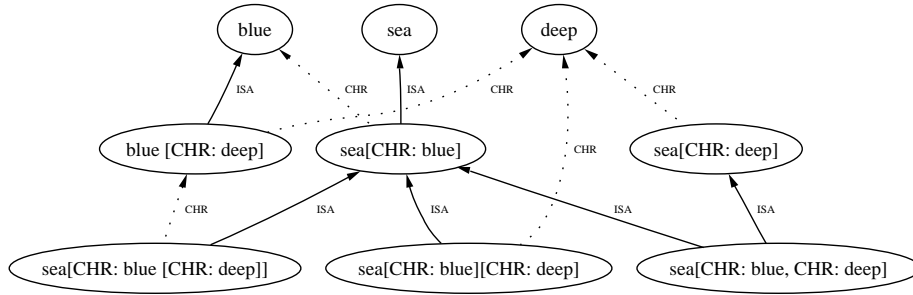


Fig. 2. A visualization of different semantic interpretations of the expression “*deep blue sea*”.

2.3 Visualizing Ontolog expressions

ONTOLOG expression can be translated into directed graphs, by analyzing the semantics of the expression. An ONTOLOG expression, O consists of a set of compound concepts C , a set of (directed) semantic relations R and a mapping $\delta: R \rightarrow C \times C$, relating compound concepts.

Figure 2 is the visualization of the expression “*deep blue sea*”, which has three different semantic interpretations:

- $sea[CHR: blue, CHR: deep]$, the sea which is blue and deep.
- $sea[CHR: blue][CHR: deep]$, the blue sea which is deep.
- $sea[CHR: blue [CHR: deep]]$, the sea whose (color) is deep blue.

In ONTOLOG we have only two different interpretations because the $sea[CHR: blue, CHR: deep]^2$ and $sea [CHR : blue] [CHR : deep]$ are equal due to the associative nature of ONTOLOG.

² The expression $sea[CHR: blue, CHR: deep]$ is equivalent to $sea \begin{bmatrix} CHR: blue \\ CHR: deep \end{bmatrix}$.

3 From ontology to similarity

In building a query evaluation principle that draws on an ontology, a key issue is of course how the ontology influence the matching of values, that is, how the different relations of the ontology may contribute to similarity. We have to 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.

We discuss firstly below how to introduce similarity based on the key ordering relation in the ontology: hyponymy relation³ as applied on atomic concepts (concepts explicitly represented in the ontology). Secondly we discuss how to extend the notion of similarity to cover – not only atomic but – general compound concepts as expressions in the language ONTOLOG. This involves the semantic relations introduced above as part of the language and as complementing the primary hyponymy relation.

3.1 Similarity on atomic concepts

In the OntoQuery project the hyponymy or concept inclusion relation plays a central role as the ordering relation that bind the ontology in a lattice. Concept inclusion intuitively imply strong similarity in the opposite direction of the inclusion (specialization), but also the direction of the inclusion (generalization) must contribute with some degree of similarity. Take as an example the small fraction of an ontology in figure 3. With reference to this ontology the atomic concept *dog* can be directly expanded to cover also *poodle* and *alsatian*.

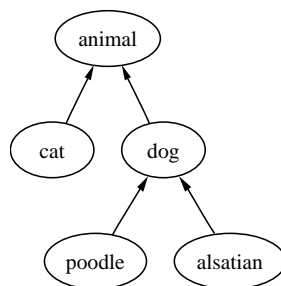


Fig. 3. Inclusion relation (*ISA*) with upwards reading, e.g. *dog ISA animal*.

This expansion respects the ontology in the sense that every instance of the extension of the expanded concept *dog* (that is, every element in the union of the extensions of *dog*, *poodle* and *alsatian*) by definition bear the relation *ISA* to *dog*. The intuition is that to a query on *dog* an answer including instances

³ Concept inclusion (*ISA*).

poodle is satisfactory (a specific answer to a general query). Since 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.

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

A concept inclusion relation can be mapped into a similarity function in accordance with the described intuition as follows.

Assume an ontology reflecting domain knowledge and comprising a partial ordering *ISA*-relation. Figure 3 shows an example. Apart from the references shown as edges a number of implicit references, for instance *poodle ISA animal*, can be derived because of the transitivity. To make "distance" influence similarity we need either to be able to distinguish explicitly stated, original references from derived or to establish a transitive reduction of the *ISA* relation. Let ISA^\sim be a stated or derived non-transitive relation such that *ISA* becomes the transitive closure of ISA^\sim . Similarity reflecting distance can then be measured from path-length in the ISA^\sim lattice. A similarity function *sim* based on distance in ISA^\sim $dist(X, Y)$ 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: If there is a path from nodes (concepts) X and Y in the hyponymy relation then it has the form

$$P = (P_1, \dots, P_n) \text{ where } P_i ISA^\sim P_{i+1} \text{ or } P_{i+1} ISA^\sim P_i \text{ for each } i \quad (1)$$

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:

$$s(P) = |\{i \mid P_i ISA^\sim P_{i+1}\}| \text{ and } g(P) = |\{i \mid P_{i+1} ISA^\sim P_i\}| \quad (2)$$

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

$$sim(X, Y) = \max_{j=1, \dots, m} \left\{ \sigma^{s(P^j)} \gamma^{g(P^j)} \right\} \quad (3)$$

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. An atomic concept T can then be expanded to a fuzzy set, including

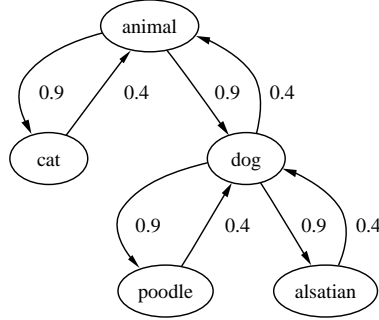


Fig. 4. The ontology transformed into a directed weighted graph, with the immediate specialization and generalization similarity being $\sigma = 0.9$ and $\gamma = 0.4$ respectively as weights. Similarity is derived as maximal (multiplicative) weighted path length, thus $sim(poodle, alsatian) = 0.4 * 0.9 = 0.36$.

T and similar values T_1, T_2, \dots, T_n as in:

$$T+ = 1/T + sim(T, T_1)/T_1 + sim(T, T_2)/T_2 + \dots + sim(T, T_n)/T_n \quad (4)$$

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:

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

3.2 General concept-similarity

The semantic relations, used in forming concepts in the ontology, indirectly contribute to similarity through subsumption. For instance *disorder*[CBY: *lack* [WRT: *vitaminD*]] is subsumed by - and thus extensionally included in - each of the more general concepts *disorder*[CBY: *lack*] and *disorder*. Thus with a definition of similarity covering atomic concepts, and in some sense reflecting the ordering concept inclusion relation, we can extend to similarity on compound concepts by a relaxation, which 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

disorder[*CBY: lack*[*WRT: vitaminD*]]

is expanded to the set

$\{\textit{disorder}, \textit{disorder}[\textit{CBY: lack}], \textit{disorder}[\textit{CBY: lack}[\textit{WRT: vitaminD}]]\}$

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

For instance a query on *disease* could be expanded to a query on similar values like:

$\textit{disease}+ = 1/\textit{disease} + \dots + 0.4/\textit{disorder} + \dots$

and a potential answer object like *disorder*[*CBY: lack*[*WRT: vitaminD*]] 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. It is essential that similarity based on subsumption expansion, exploits that for instance the degree to which $c[r_1: c_1]$ is matching $c[r_1: c_1[r_2: c_2]]$ is higher than the degree for c with no attributes is matching $c[r_1: c_1[r_2: c_2]]$. Approaches to aggregation that can be tailored to obtain these properties, based on order weighted averaging[11] and capturing nested structuring[12], are described in [1, 2].

An alternative to the above described subsumption expansion could be to generalize the principle of weighted path similarity as described in the previous subsection for the ISA-relation. 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 *disorder*[*CBY: lack*] is conceptually not some kind of a *lack*. On the other hand it would be reasonable to claim that *disorder*[*CBY: lack*] in a broad sense has something to do with (and thus has similarities to) *lack*. 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 NL phrases, which is underlying any strictly word-based approach, does not lead to serious failure because the correlation that appears is not dominating.

This could of course be an argument for not looking at compound concepts at all, but rather these considerations points in the direction of redrawing some of the importance of correlation in NL phrases when developing similarity measures.

Consider figure 5. The solid edges are *ISA* references and the broken are references by other semantic relations – in this example *CBY* and *WRT* are in use. When ignoring the broken edges, the graph shown is a subgraph of the key lattice of the ontology (where *ISA* is the ordering relation). Each compound concept has broken edges to its attribution concept. The spelling out of the full compound concept expression as the label of a node is redundant (the concept expression can be derived from the connecting edges).

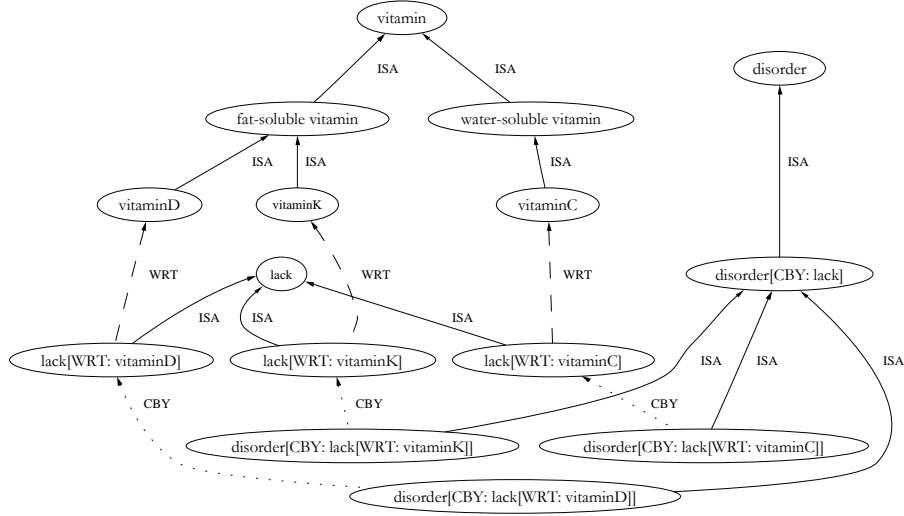


Fig. 5. A graph covering part of the ontology with semantic relations as paths.

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 does not apply to semantic relations and seemingly there is no obvious way to differentiate based on direction at all. Thus one approach in 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:

$$r^j(P) = |\{i \mid P_i R^j P_{i+1}\}| \quad (5)$$

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

$$sim(X, Y) = \max_{i=1, \dots, m} \left\{ \sigma^{s(P^i)} \gamma^{g(P^i)} r^1(P^i) \dots \rho_k^{r^k(P^i)} \right\} \quad (6)$$

Take as an example the ontology of figure 5. Here two semantic relations WRT 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} .

Figure 6 shows the transformed graph with the attached similarity factors as weights, again assuming that the degree to which a concept Y is similar to a given concept X can be derived as the maximum of the products of edge weights over the set of paths connecting X and Y .

The attached weights are in the example assigned values in a rather ad hoc manner. Such assignment in practice needs a careful effort by domain experts. Furthermore the similarity principle in general needs to be verified empirically.

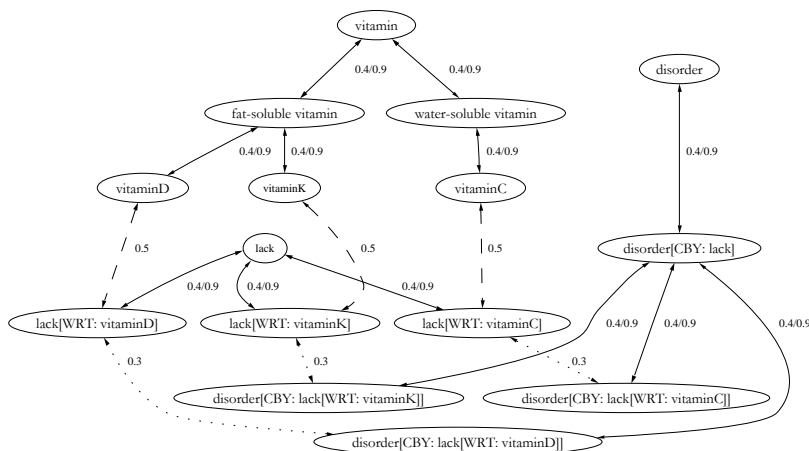


Fig. 6. The ontology of figure 5 transformed into a directional weighted graph with the similarity factors for specialization: $\sigma = 0.9$, for generalization: $\gamma = 0.4$, for CBY : $\rho_{CBY} = 0.3$ and for WRT : $\rho_{WRT} = 0.5$. The upwards and downwards edges from figure 4 describing generalization and specialization respectively, are here merged into one edge pointing in both directions. The weights are given as a pair generalization/specialization, respectively, as label on the edges. Similarity is derived as maximal (multiplicative) weighted path length, thus $sim(disorder[CBY: lack[WRT: vitaminD]], lack) = 0.3 * 0.4 = 0.12$.

4 Conclusion

We have described three different principles for measuring similarity between both atomic and compound concepts, all of which incorporate meta knowledge.

- Similarity between atomic concepts based on distance in the ordering relation of the ontology, concept inclusion (ISA).
- Similarity between general compound concepts based on subsumption expansion.
- Similarity between general compound concepts based on distance in the ontology, but aggregated with distances in the semantic relations.

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.

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. 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.

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: *OntoQuery: Ontology-based Querying of Texts*, AAAI 2002 Spring Symposium, Stanford, California, 2002.
- [5] 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.
- [6] 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, to be republished at www.ontoquery.dk.
- [7] Meadow, C.T., Boyce, B.R., Kraft, D.H.: *Text information retrieval systems*, second edition, Academic Press, 2000.
- [8] Nistrup, B. Madsen and B. Sandford Pedersen and H. Erdman Thomsen: Semantic Relations in Content-based Querying Systems: a Research Presentation from the OntoQuery Project, in [6].
- [9] Nilsson, J. Fischer: A Logico-algebraic Framework for Ontologies ONTOLOG, in [6].
- [10] ONTOQUERY project net site: www.ontoquery.dk
- [11] Yager, R.R.: On ordered weighted averaging aggregation operators in multicriteria decision making, in *IEEE Transactions on Systems, Man and Cybernetics*, vol 18, 1998.
- [12] Yager, R.R.: A hierarchical document retrieval language, in *Information Retrieval* vol 3, Issue 4, Kluwer Academic Publishers pp. 357-377, 2000.