

From Ontology over Similarity to Query Evaluation

Troels Andreasen, Henrik Bulskov, Rasmus Knappe

*Department of Computer Science, Roskilde University, P.O. Box, DK-4000,
Roskilde, Denmark*

Abstract

The main focus of this paper concerns the measuring and application of similarity in a content-based information retrieval environment. Documents or information base objects are assumed to be indexed by concepts, that is, expressions in a simple description language ONTOLOG. Descriptions refer to a generative ontology, assumed to reflect a domain with specific words and concepts, and queries are to be evaluated by application of similarity that in turn reflects the structure and relations in the ontology. Similarity is derived using the notion of a similarity graph and functions grading similar values to given values – allowing numerical computation rather than symbolic reasoning during query evaluation are proposed. The application of similarity in this environment raises important aspects concerning scalability as discussed in this paper. Finally a suggested principle of evaluation is given and its scalability is analyzed.

Key words: Fuzzy Similarity, Flexible Querying, Ontology

1 Introduction

The objective of this paper is to devise and discuss principles for evaluation of similarity measures that utilizes knowledge from an ontology to obtain better and closer 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 base.

Email addresses: troels@ruc.dk (Troels Andreasen), bulskov@ruc.dk (Henrik Bulskov), knappe@ruc.dk (Rasmus Knappe).

The ontology plays its role behind the scenes, it defines and relates the concepts that are the basis for comparing queries and answers. Our claim is that when the ontology is no longer the primary base in focus, a more restrictive language with less expressive power is more suited in the present context. The main argument for this is that we can do with an incremental volume of knowledge represented in the ontology. Even very small fragments from a domain, such as a few related concepts, makes sense as an ontology if answers to queries can be improved by this. There is no need at all to insist on completeness on the coverage of a domain or a subdomain.

We consider a generative framework where an ontology in combination with a concept language defines a set of well-formed concepts. Well-formed concepts are assumed to be the basis for an indexing of the information base in the sense that these concepts appear as descriptors attached to objects in the base. Descriptors are combined in descriptions to express the semantic understanding of fragments of text, i.e. sentences.

This conceptual indexing is sought done by extraction of concepts from the information base by simple partial linguistic analysis. The lack of completeness is not decisive because every little step from indexing by terms towards indexing with concepts is a step towards conceptual indexing.

In this context, one of the major problems is to determine the similarity between the semantic elements. It is no longer only simple match of keywords in the text objects, but also the meaning of them, we have to take into consideration when we calculate the similarity between queries and objects in the base.

The devised similarity measure forms the basis for an evaluation principle that on the one hand incorporate the knowledge in the ontology and the semantics of concepts used for description of queries and indexing of the information base, and on the other hand is realistic as an evaluation principle for large scale information systems.

2 Concept Language

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

The concept language in focus here, ONTOLOG (Fischer Nilsson, 2001), defines a set of semantic relations which can be used for “attribution” (feature-attachment) to form compound concepts. The suitable number of available relations may vary with different domains, but among the more general relations that probably will be present in most domain modelings are WRT (With-respect-to), CHR (Characterized-by), CBY (Caused-by), TMP (Temporal), LOC (Location). Expressions in ONTOLOG are descriptions of concepts situated in an ontology formed by an algebraic lattice with concept inclusion (ISA) as the ordering relation (Knappe et al., 2003). Attribution of concepts, i.e combining atomic concepts into compound concepts by attaching attributes, can be written as a feature structures. Simple attribution of a concept c_1 with relation r and a concept c_2 is denoted $c_1[r: c_2]$. Take as an example the sentence: “*the black dog is making noise*” which can be translated into this semantic expression $noise[CBY: dog[CHR: black]]$.

Descriptions of text expressed in this language describe semantics and goes beyond simple keyword descriptions. 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 Measuring Similarity

Obviously there is no such thing as uniqueness as related to general proximity in knowledge. Moreover from just considering the potential dimensions involved it should be apparent that reasoning on knowledge can be a task of almost arbitrary complexity. So before going further into a discussion on how to measure similarity we emphasize the following. Firstly we cannot expect to find a universal measure that can be used independently of the knowledge represented and the domain in question. Rather the modest and pragmatic aim in developing measures of “conceptual similarity” is to demonstrate usefulness according to a set of preferred properties. Secondly it is essential to take computational complexity in deriving similarity into account – especially as any kind of query environment requires this. In general such environments should be capable of handling huge amounts of information base objects.

Similar concepts are concepts that have much in common. To approximate this vague notion we derive conceptual similarity using the notion of a “similarity graph”.

A similarity graph (Knappe et al., 2003) is a subpart of the ontology represented as a graph with a subset of concepts as nodes and relations connecting these as edges. This narrowing of relevant concepts enables a move toward a reduced complexity by transforming formal conceptual reasoning into numer-

ical concept similarity computation. 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.

Fig. 1 shows an example of a similarity graph covering two terms $cat[CHR: black]$ and $poodle[CHR: black]$

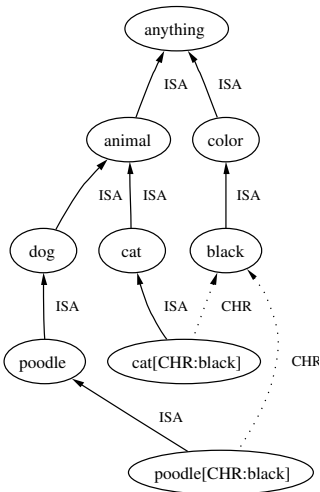


Fig. 1. An example of a similarity graph for the concepts $cat[chr: black]$ and $poodle[chr: black]$

3.1 Similarity

As already mentioned above we aim to derive similarity from a vaguely defined notion of how much concepts have in common. Our objective is thus to derive a function $sim(x, y)$ that measure degree of similarity proportional to how much the concepts x and y share or how close they are. Without loss of generality we assume that the function maps concepts into the unit interval:

$$sim(x, y) : C \times C \rightarrow [0, 1]$$

where C is the set of well-formed concepts and where $sim(x, y)$ measure the degree to which y is similar to x . The extreme values $sim(x, y) = 0$ means not similar and $sim(x, y) = 1$ means fully similar. The latter may only be the case when $x = y$.

When restricting to similarity graphs one obvious approach is to reflect what connects the concepts x and y . As discussed in previous work (Bulskov et al., 2002) this can be done by considering the shortest path connecting the concepts x and y . As raised in (Knappe et al., 2003; Andreasen et al., 2003)

the shortest path approach to similarity lacks the influence of an important aspect that has to do with multiple connections between concepts. 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 $\alpha(x)$ (Knappe et al., 2003) as the set of nodes (upwards) reachable from x , we have $\alpha(x) \cap \alpha(y)$ as the reachable nodes shared by x and y , which thus obviously is an indication of what’s 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 $\alpha(y)$ thus 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)

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

There is no question about that we by similarity graphs and functions as the above only obtain a very coarse-grained approximation of whatever 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, if only the measuring 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 a given functions accordance with the semantics of the ontology and use these to guide the choice of function.

First of all it is important to notice that the similarity graph principle unifies the concept inclusion relation with the semantic relations used in attribution. We still consider upwards and downwards in the unified graph as generalization and specialization respectively, but it is important to notice that this is no longer strictly subsumption based. We thus take not only *cat* but also *black* to be “generalizations” of *cat*[*CHR* : *black*] and even *cat*[*CHR* : *black*] to be a generalization of *accident*[*CBY* : *cat*[*CHR* : *black*]].

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

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

The intuition being that for instance a “cat” satisfies the intention of an “animal” while an “animal” (that could be of any kind) not necessarily satisfies the intention of a “cat”. From this property alone we can eliminate the first alternative similarity function (a) above. The consequence of insisting on this property is namely that the similarity function cannot be symmetrical, which (a) obviously is. In fig. 2, for instance, the (a) alternative similarity function gives $sim(D, E) = sim(E, D) = \frac{2}{5}$, while we should have $sim(D, E) < sim(E, D)$ according to the generalization cost property.

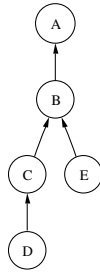


Fig. 2. *Generalization cost property* implies $sim(D, E) < sim(E, D)$ and *Specialization cost property* implies $sim(C, E) > sim(D, E)$

Now consider alternative (c). We get in fig. 2 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 obey the property is (b). With the example in fig. 2 we get $sim(D, E) = \frac{2}{4}$ and $sim(E, D) = \frac{2}{3}$

A second property that tends to appear more as optional since 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 nodes are more specific.

The intuition for this property is that the similarity between for instance siblings on low levels in the ontology as “alsatian” and “poodle” should be higher than the similarity between siblings close to the top as “Physical” and “Abstract”. This idea corresponds to the notion of information content described in (Resnik, 1998).

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

A third property that similarly cannot be claimed to be semantically implied

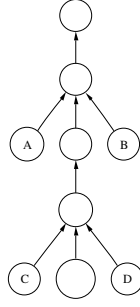


Fig. 3. *Specificity cost property*: implies that $sim(C,D) > sim(A,B)$.

is the following.

Specialization cost property – further specialization implies reduced similarity.

As support of the intuition for this property consider again fig. 2. The (b) similarity function does obviously not satisfy this property since 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 to consider alternative similarity functions that are influenced by both specializations and generalizations (as the function (a) above is), but still not violates the only ultimate property above; 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 the following:

(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 is in favor of 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. (b) is just a special case of (d) with $\rho = 1$, the parameter ρ allows to tailor the similarity function, and can thereby comply with the generalization property.

Example

As illustration of how (b) and (d) differs consider the subontology in fig. 1. The similarities for *poodle* respectively *poodle[CHR : black]* and the other concepts included in the subontology (i.e similarity graph) are, when collected in fuzzy subsets of similar concepts (with $similar(x) = \Sigma sim(x, y)/y$) the following:

For (b) we get

$$\begin{aligned} \text{similar}(\text{poodle}) = & \\ & 1.00/\text{poodle} + 1/\text{poodle}[\text{CHR} : \text{black}] + 0,75/\text{dog} + 0,5/\text{animal} + 0,5/\text{cat} + \\ & 0,5/\text{cat}[\text{CHR} : \text{black}] \end{aligned}$$

$$\begin{aligned} \text{similar}(\text{poodle}[\text{CHR} : \text{black}]) = & \\ & 1.00/\text{poodle}[\text{CHR} : \text{black}] + 0,57/\text{poodle} + 0,57/\text{cat}[\text{CHR} : \text{black}] + 0,43/\text{dog} + \\ & 0,43/\text{black} + 0,29/\text{animal} + 0,29/\text{cat} + 0,29/\text{color} \end{aligned}$$

and (d) with $\rho = \frac{4}{5}$ leads to

$$\begin{aligned} \text{similar}(\text{poodle}) = & \\ & 1.00/\text{poodle} + 0,91/\text{poodle}[\text{CHR} : \text{black}] + 0,80/\text{dog} + 0,60/\text{animal} + 0,53/\text{cat} + \\ & 0,47/\text{cat}[\text{CHR} : \text{black}] + 0,30/\text{color} \end{aligned}$$

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

4 Evaluation Principles

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.

In addition to the problem of finding a useful measuring principle a challenge is to devise a principle of similarity-based evaluation that is realistic in connection with query processing.

To this end the principle of similarity expansion is an obvious improvement. Instead of calculating similarities in connection with every matching of two values during evaluation, one of these can be expanded and similarity matching becomes a matter of value to set comparison. As indicated through the example above we can introduce similar values by expanding a crisp value into a fuzzy set including also similar values.

For query processing we can choose either to expand every index term connected to objects in the information base or every term in the query. In the approach presented here we choose the latter, which obviously involves decisively less storage space.

In principle we need for a given concept to compare and derive similarity for every other concept in the database. Now since the ontology is generative the set of well-formed concepts is not finite. However for query processing we need only to compare concepts in use (used in the query or used in the index

Table 1

Similarity $sim(c_i, c_j)$ as defined for concepts C included in figure 1. The numbers used for column labels correspond with the numbers written in parenthesis in the row labels.

	1	2	3	4	5	6	7	8	9
(1) anything	1,00	0,90	0,90	0,87	0,87	0,87	0,85	0,83	0,83
(2) animal	0,60	1,00	0,50	0,93	0,93	0,47	0,90	0,87	0,86
(3) color	0,60	0,50	1,00	0,47	0,47	0,93	0,45	0,87	0,86
(4) dog	0,47	0,73	0,37	1,00	0,67	0,33	0,95	0,60	0,89
(5) cat	0,47	0,73	0,37	0,67	1,00	0,33	0,63	0,90	0,59
(6) black	0,47	0,37	0,73	0,33	0,33	1,00	0,32	0,90	0,89
(7) poodle	0,40	0,60	0,30	0,80	0,53	0,27	1,00	0,47	0,91
(8) cat[CHR:BLACK]	0,33	0,47	0,47	0,40	0,60	0,60	0,37	1,00	0,65
(9) poodle[CHR:BLACK]	0,31	0,43	0,43	0,54	0,36	0,54	0,66	0,59	1,00

in the database), so we can reduce to the finite number CN of terms that are either atomic terms in the ontology or compound terms in the database. Applying the chosen similarity function we need in principle to derive the similarity between a given concept and all concepts $C = \{c_1, \dots, c_{CN}\}$ in the ontology/database, thus to expand a concept c to

$$similar(c) = \sum_{i=1}^{CN} sim(c, c_i)/c$$

This can be viewed as a 2-dimensional matrix having size n^2 where the value of the (i, j) 'th entry is the similarity between the i 'th and j 'th concept.

Table 1 shows similarity by definition (d) with $sim(c_i, c_j)$ as the value of the (i, j) 'th entry. Here $C = \{anything, animal, color, dog, cat, black, poodle, cat[CHR:BLACK], poodle[CHR:BLACK]\}$.

In order not to expand with all concepts in the ontology the expansion can be restricted by a given threshold β used as a cut in the fuzzy set corresponding to the expansion. The expansion for *poodle* with $\beta = 4/5$ is $similar(poodle) = 1.00/poodle + 0.80/dog + 0.91/poodle[CHR:BLACK]$.

In general we define the expansion of a query Q as the set of expanded query terms, thus with $Q = \{c_1, \dots, c_n\}$ the expansion is

$$E(Q) = \{similar(c_1), \dots, similar(c_n)\},$$

thus with query $Q = \{poodle, black\}$ and threshold $\beta = 4/5$ we get the following expanded query:

$$E(\{poodle, black\}) = \left\{ \begin{array}{l} 1.00/poodle + 0.80/dog + 0.91/poodle[CHR : black], \\ 1.00/black + 0.90/cat[CHR : black] + 0.89/poodle[CHR : black] \end{array} \right\}$$

The expansion principle used for atomic concepts can be generalized to cover all concepts by performing term decomposition on all compound concepts. The term decomposition (as defined in (Andreasen et al., 2003)) thereby replaces symbolic reasoning over the ontology to numerical computation by expanding a given compound concept with the term decomposition of the concept. This decomposition is performed for both the terms in the query and for the well-formed concepts indexing the objects in the information base.

The evaluation principle is defined as follows. We define a similarity function that calculates the similarity between a query $Q = \{Q_1, \dots, Q_n\}$ and an object O in the information base. Every base object O is assumed to be indexed by a set of well-formed concepts $\{O_1, \dots, O_k\}$.

The evaluation of $similarity(Q, O)$ requires a nested aggregation principle, due to the fact that we have both query expansion and an indexing principle where objects in the information base are indexed by the decomposition of their well-formed concepts. The generalization of order weighted averaging aggregation (Yager, 1988), which is presented in (Yager, 2000) as “hierarchical aggregation” can be used here. We shall not go into details on this general principle but present below a simplified nested aggregation, which is a special case, based on arithmetic average.

Let $Q' = E(Q) = \{similar(Q_1), \dots, similar(Q_n)\} =$

$\{w_{11}/q_{11} + \dots + w_{1m_1}/q_{1m_1}, \dots, w_{n1}/q_{n1} + \dots + w_{nm_n}/q_{nm_n}\}$ be the expansion of Q and $O' = \{O'_1, \dots, O'_k\} = \{\{o_{11}, \dots, o_{1l_1}\}, \dots, \{o_{k1}, \dots, o_{kl_k}\}\}$ be the decomposition of all concepts $\{O_1, \dots, O_k\}$ indexing an object O . Then the overall similarity between the expanded query Q' and a decomposed object O' can be done by an aggregation over the similarity between each term $\{Q'_1, \dots, Q'_n\}$ in the query and the object O' .

$$similarity(Q, O) = \frac{1}{n} \sum_{i=1}^n sim1(Q'_i, O') \quad (1)$$

This is done by calculating the maximal similarity between each term in the query Q'_i and each well-formed concept $\{O'_1, \dots, O'_k\}$ in the object O' .

$$sim1(Q'_i, O') = \max_{j=1, \dots, k} (sim2(Q'_i, O'_j)) \quad (2)$$

In order to do so the similarity between the fuzzy set Q'_i and the well-formed concepts in the decomposition of O'_j is calculated. This can be done based on an aggregation over the membership grade of each concept in O'_j in the fuzzy set Q'_i . For this purpose the membership function $\mu_{Q'_i}(x)$ for Q'_i is used, giving the degree of membership for x to the fuzzy set Q'_i .

$$sim2(Q'_i, O'_j) = \frac{1}{n} \sum_{h=1}^{l_j} \mu_{Q'_i}(O'_{kh}) \quad (3)$$

The overall complexity of the evaluation principle presented here is not insuperable. The hardest part, the calculation of the similarity matrix, can be preprocessed and will therefore not influence the complexity of the query evaluation. Maintenance of the similarity matrix can be done incrementally and therefore without the need for recalculating similarity between existing concepts when adding compound concepts to the ontology.

5 Conclusion

We have described a notion of generative ontology and a principle for measuring similarity that reflect the generative nature of the ontology.

Similarity graphs as the basis for similarity measures exploiting shared nodes are introduced and seems to indicate a usable theoretical and practical foundation for design of conceptual similarity measures.

The evaluation principle described aims to incorporate the knowledge represented in the ontology and by the well-formed concepts indexing the objects in the information base. In order to evaluate the scalability of the proposed evaluation principle a prototype system is subject to ongoing implementation for empirical studies.

Acknowledgments

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

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.

References

- 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
- 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.
- Andreasen, T., Bulskov, H. and Knappe, R.: On ontology-based querying, pp. 53-59 in Heiner Stuckenschmidt (Eds.): 18th International Joint Conference on Artificial Intelligence, Ontologies and Distributed Systems, IJCAI 2003, Acapulco, Mexico, August 9, 2003, Workshop Program
- Andreasen, T., Bukskov, H., and Knappe, R.: Similarity from Conceptual Relations, pp. 179-184 in Ellen Walker (Eds.): 22nd International Conference of the North American Fuzzy Information Processing Society, NAFIPS 2003, Chicago, Illinois USA, July 24-26, 2003, Proceedings
- 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
- Knappe, R., Bulskov, H. and Andreasen, T.: Similarity Graphs, LNAI 2871, pp. 668-672 in N. Zhong, Z.W. Ras, S. Tsumoto, E. Suzuki (Eds.): 14th International Symposium on Methodologies for Intelligent Systems, ISMIS 2003, Maebashi, Japan, October 28-31, 2003, Proceedings
- Knappe, R., Bulskov, H. and Andreasen, T.: On Similarity Measures for Content-based Querying, pp. 400-403 in O. Kaynak et. al. (Eds.): 10th International Fuzzy Systems Association World Congress, IFSA 2003, Istanbul, Turkey, June 29-July 2, 2003, Proceedings
- 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.
- Resnik, P.: Semantic similarity in a taxonomy: an information-based measure and its application to problem of ambiguity in natural language. *J. Art. Int. Res.*, 1998.
- Yager, R.R.: On ordered weighted averaging aggregation operators in multi-criteria decision making, in *IEEE Transactions on Systems, Man and Cybernetics*, vol 18, 1988.
- Yager, R.R.: A hierarchical document retrieval language, in *Information Retrieval* vol 3, Issue 4, Kluwer Academic Publishers pp. 357-377, 2000.