A Survey of Semantic Similarity Methods for Ontology based Information Retrieval

K.Saruladha,Senior Lecturer,
Department of Computer Science & Engg.
Pondciherry Engineering College,
Pondicherry, India.
charusanthaprasad@yahoo.com

Dr.G.Aghila
Department of Computer Science,
Pondciherry
University,Pondicherry,India.
aghilaa@yahoo.com

Sajina Raj, Post Graduate Student,
Department of Computer Science & Engg.
Pondicherry Engineering College,
Pondicherry, India
sajinaraj@pec.edu

***Abstract*— This paper discusses the various approaches used for identifying semantically similar concepts in an ontology. The purpose of this survey is to explore how these similarity computation methods could assist in ontology based query expansion. This query expansion method based on the similarity function is expected to improve the retrieval effectiveness of the ontology based Information retrieval models. Various similarity computation methods fall under three categories: Edge counting, information content and node based counting. The limitations of each of these approaches have been discussed in this paper.**

Keywords-Ontology, similarity method, information retrieval, concept ual similarity, taxonomy, corpus based

1 . INTRODUCTION

The goal of Information retrieval process is to retrieve Information relevant to a given request. The aim is to retrieve all the relevant information eliminating the non-relevant information. An information retrieval system comprises of representation, semantic similarity matching function and Query. Representation comprises the abstract description of documents in the system. The semantic similarity matching function defines how to compare query requests to the stored descriptions in the representation.

The percentage of relevant information we get mainly depends on the semantic similarity matching function we used. So far, there are several semantic similarity methods used which have certain limitations despite the advantages. No one method replaces all the semantic similarity methods. When a new information retrieval system is going to be build, several questions arises related to the semantic similarity matching function to be used. In the last few decades, the number of semantic similarity methods developed is high.

This paper discusses the survey of different similarity measuring methods used to compare and find very similar concepts of an ontology. Section II, a set of basic intuitive properties are defined to which the compatibility of similarity measures in information is preferable. Section III discusses various approaches used for similarity computation. In Section IV a comparison of various similarity computation methods based on the experiments conducted to evaluate them is discussed. Finally how do similarity computations and human judgments are correlating is tabulated.

2. ONTOLOGY SIMILARITY

In this section, a set of intuitive and qualitative properties that a similarity method should adhere to is discussed.

- *Basic Properties*

  Any similarity measure must be compatible with the basic properties as they express the exact notion of property.
  - *CommonalityProperty*
  - *Difference Property*
  - *Identity Property*
- *Retrieval Specific Properties*

  The similarity measure cannot be symmetrical in case of ontology based information retrieval context. The similarity is directly proportional to specialization and inversely proportional to generalization.
  - Generalization Property
- *Structure Specific Properties*

  The distance represented by an edge should be reduced with an increasing depth.
  - *Depth Property*
  - *Multiple Paths Property*

3. APPROACHES USED FOR SIMILARITY COMPUTATION

In this section, we discuss about various similarity methods. The similarity methods are

- *Path Length Approaches*
- *Depth-Relative Approaches*
- *Corpus-based Approaches*
- *Multiple-Paths Approaches*

3.1 *Path Length Approach*

The shortest path length and the weighted shortest path are the two taxonomy based approaches for measuring similarity through inclusion relation.

Shortest Path Length

978-0-7695-3977-5/10 $26.00 © 2010 IEEE
DOI 10.1109/ICMLC.2010.63

A simple way to measure semantic similarity in a taxonomy is to evaluate the distance between the nodes corresponding to the items being compared. The shorter distance results in high similarity.

In Rada et al. [1989][1][14], shortest path length approach is followed assuming that the number of edges between terms in a taxonomy is a measure of conceptual distance between concepts.

distRada(*ci; cj*) = minimal number of edges in a path from *ci* to *cj*

This method yields good results. since the paths are restricted to ISA relation, the path lengths corresponds to conceptual distance. Moreover, the experiment has been conducted for specific domain ensuring the hierarchical homogeneity.

The drawback with this approach is that, it is compatible only with commonality and difference properties and not with identity property.

3.1.2. Weighted Shortest Path Length

This is another simple edge-counting approach. In this method, weights are assigned to edges. In brief, weighted shortest path measure is a generalization of the shortest path length. Obviously it supports commonality and difference properties.

Given a path P=(p_1,…..p_n), set s(P) to the number of specializations and g(P) to the number of generalizations along the path P as follows:

$$s(P)= |\{i\backslash p_i \text{ ISA } p_{i+1}\}| \quad (1)$$

$$g(P)=|\{i|P_{i+1} \text{ ISA } p_i\}| \quad (2)$$

If p_1,……p_m are all paths connecting x and y, then the degree to which y is similar To x can be defined as follows:

$$simWSP(x,y)=\max\{ {}^{s(pj)} \quad {}^{s(pj)}\} (3)$$

$$j=1,\ldots.m$$

The similarity between two concepts x and y, sim(x,y) WSP(weighted Shortest Path) is calculated as the maximal product of weights along the paths between *x* and *y*. Similarity can be derived as the products of weights on the paths.

$$\alpha \in [0,1] \text{ and } \gamma \in [0,1] \max^{s(pj)}$$

$$g(Pj \quad s(P^j) \text{ and } g(P^j) = 0$$

Hence the weighted shortest path length overcomes the limitations of shortest path length wherein the measure is based on generalization property and achieves identity property.

3.2 Depth-Relative Approaches

Even though the edge counting method is simple, it limits the representation of uniform distances on the edges in the taxonomy. This approach supports structure specific property as the distance represented by an edge should be reduced with an increasing depth.

3.2.1 Depth Relative Scaling

In his depth-relative scaling approach Sussna[1993][2] defines two edges representing inverse relations for each edge in a taxonomy. The weight attached to each relation *r* is a value in the range [min*r*; max*r*]. The point in the range for a relation *r* from concept *c*1 to *c*2 depends on the number *nr* of edges of the same type, leaving *c*1, which is denoted as the type specifc fanout factor:

$$W(c_{1\rightarrow r}\, c_2)=\max_r\text{-}\{\max_r\text{-}\min_r/n_r(c_1)\}$$

The two inverse weights are averaged and scaled by depth *d* of the edge in the overall taxonomy. The distance between adjacent nodes *c*1 and *c*2 are computed as:

$$\text{dist}_{\text{sussna}}(c_1,c_2)=w(c_{1\rightarrow r}\, c_2)+ (c_{1\rightarrow r'}\, c_2)/2d \quad (4)$$

where *r* is the relation that holds between *c*1 and *c*2, and *r'* is its inverse. The semantic distance between two arbitrary concepts *c*1 and *c*2 is computed as the sum of distances between the pairs of adjacent concepts along the shortest path connecting *c*1 and *c*2.

3.2.2 Conceptual Similarity

Wu and Palmer [1994][3], propose a measure of semantic similarity on the semantic representation of verbs in computer systems and its impact on lexical selection problems in machine translation. Wu and Palmer define *conceptual similarity* between a pair of concepts *c*1 and *c*2 as:

$$\text{Sim}_{\text{wu\&palmer}}(c_1,c_2)= \quad 2*N3/N \quad (5)$$

Where N1 is the number of nodes on the path from c1 to a concept c3. , denoting the least upper bound of both *c*1 and *c*2. *N*2 is the number of nodes on a path from *c*2 to *c*3. *N*3 is the number of nodes from *c*3 to the most general concept.

3.2.3 Normalised Path Length

Leacock and Chodorow [1998][4], proposed an approach for measuring semantic similarity as the shortest path using is a hierarchies for nouns in WordNet. This measure determines the semantic similarity between two synsets (concepts) by finding the shortest path and by scaling using the depth of the taxonomy:

$$\text{Sim}_{\text{Leacock\&Chaodorow}}(c_1,c_2)= -\log(N_p(c_1,c_2)/2D) \quad (6)$$

Np (*c*1,*c*2) denotes the shortest path between the synsets (measured in nodes), and *D* is the maximum depth of the taxonomy.

3.3 Corpus-based Approach

The knowledge disclosed by the corpus analysis is used to intensify the information already present in the ontologies or taxonomies. In this method, presents three approaches that incorporate corpus analysis as an additional, and qualitatively different knowledge source.

3.3.1 Information Content

In this method rather than counting edges in the shortest path, they select the maximum information content of the least upper bound between two concepts. Resnik [1999] [5], argued that a widely acknowledged problem with edge-counting approaches was that they typically rely on the notion that edges represent uniform distances. According to Resnik's measure, information content, uses knowledge

from a corpus about the use of senses to express non-uniform distances.

Let C denote the set of concepts in a taxonomy that permits multiple inheritance and associates with each concept c 2 C, the probability $p(c)$ of encountering an instance of concept c. For a pair of concepts $c1$ and $c2$, their similarity can be defined as:

$$Sim_{Resnik} = \max_{c \in S(c1,c2)} [-\log p(c)] \quad (7)$$

Where,

$S(c1,c2)$: Set of least upper bounds in the taxonomy of $c1$ and c2

$p(c)$:Monotonically non-decreasing as one moves up in the taxonomy,

$p(c1) \leq p(c2)$, if c1 is a c2.

The similarity between the two words w1 and w2 can be computed as:

$$wsim_{Resnik}(w1, w2) = \max_{c1 \in s(w1), c2 \in s(w2)} [sim \quad (8)$$

Where,

$s(wi)$: Set of possible senses for the word wi.

Resnik describes an implementation based on *information content* using WordNet's [Miller, 1990][6], taxonomy of noun concepts [1999]. The information content of each concept is calculated using noun frequencies

$$Freq(c) = \sum_{n \in words(c)}$$

Where,

words(*c*): Set of words whose senses are subsumed by concept c.

$$(c) = freq(c)/N$$

where ,*N:* is the total number of nouns.

The major drawback of the information content approach is that they fail to comply with the generalization property, due to symmetry.

3.3.2 Jiang and Conrath's Approach(Hybrid Method)

Jiang and Conrath [1997][7]proposed a method to synthesize edge-counting methods and information content into a combined model by adding the information content as a corrective factor.

The edge weight between a child concept *cc* and a parent concept *cp* can be calculated by considering factors such as local density in the taxonomy, node depth, and link type as,

$$Wt(c_c,c_p) = \left(\beta + (1-\beta)\frac{\bar{E}}{E(cp)}\right)\left(\frac{d(cp)+1}{d(cp)}\right)^{\alpha} LS(Cc, Cp)T(Cc,C \quad (9)$$

Where,

$d(cp)$: Depth of the concept cp in the taxonomy,

$E(cp)$:Number of children of cp (the local density)

(^1E) : Average density in the entire taxonomy,

$LS(cc, cp)$: Strength of the edge between cc and cp,

$T(cc,cp)$: Edge relation/type factor

Jiang and Conrath then defined the semantic distance between two nodes as the summation of edge weights along the shortest path between them [J. Jiang, 1997]:

$$dist_{jiang\&conrath}(C_1,C_2) = \sum_{ca \in \{path(c1,c2) - LSuper(c1,c2)\}} wt(c, pare \quad (10)$$

Where,

$path(c1, c2)$: the set of all nodes along the shortest path between c1 and c2

$parent(c)$: is the parent node of c

$LSuper(c1, c2)$: is the lowest superordinate (least upper bound) on the path between $c1$ *and c2*..

3.3.3 Lin's Universal Similarity Measure

Lin [1997; 1998][8][13] defines a measure of similarity claimed to be both universally applicable to arbitrary objects and theoretically justified. He achieved generality from a set of assumptions.

3.4 Multiple-Paths Approaches

This approach solves the problem with single path approach.

3.4.1 Medium-Strong Relations

Hirst and St-Onge [Hirst and St-Onge, 1998; St-Onge, 1995] [9][15], distinguishes the nouns in the Wordnet as extra-strong, strong and medium-strong relations. The extra-strong relation is only between a word and its literal repetition.

A strong relation between two words is described inHirst and St-Onge [1998]. The longer the path and the more changes in direction, the lower the weight. The *medium-strong relation* is basically a shortest path length measure and thus does not comply with the multiple-path property; hence even through it introduces both taxonomic and semantic relations, it is still restricted to only one path. It does not comply with either the generalization or depth properties, but obviously obeys the basic properties as it is a shortest path length measure.

3.4.2 Generalised Weighted Shortest Path

The principle of weighted path similarity can be generalized by introducing similarity factors for the semantic relations. However, there does not seem to be an obvious way to differentiate based on direction. Thus, we can generalize simply by introducing a single similarity factor and simplify to bidirectional edges. This method solves the symmetry problem by introducing weighted edges.

3.4.3 Shared Nodes

This approach overcomes the limitation of single path length approach. Multiple paths are considered for measuring the similarity between concepts..

The definitions of term decomposition used are:

$$r(c) = \{c\} \cup \{x | c \leq xL, \ldots, rx | \forall c \leq yL, \ldots, rx\}, x \in L, y \in L, r \in R$$

and transitive closure of a set of concepts with respect to <=

$$\omega(C') = \{x \mid x \in C \vee y \in$$

(11)

With $(x))$ as the set of nodes (upwards) reachable from x in an instantiated Ontology. Similarity function for shared weighted node by using equation 12.

$$\text{sim}_{\text{sharednodes}}(x,y) = \rho\frac{|x(x) \cap x(y)|}{|x(y)|} + (1 - \quad (12)$$

Table 1 subset of noun pairs used by niller and charles experiments

Word1	Word2	Replica	R&G	M&C
Car	Automobile	3.82	3.92	3.92
Gem	Jewel	3.86	3.84	3.94
Journey	Voyage	3.58	3.54	3.58
Boy	Lad	3.10	3.76	3.84
Coast	Shore	3.38	3.70	3.60
Asylum	madhouse	2.14	3.61	3.04
Magician	Wizard	3.68	3.50	3.21
Midday	Noon	3.45	3.42	3.94
Furnace	Stove	2.60	3.11	3.11
Food	Fruit	2.87	3.08	2.69
Bird	Cock	2.62	3.05	2.63
Bird	Crane	2.08	2.97	2.63
Tool	implement	1.70	2.95	3.66
Brother	Monk	2.38	2.82	2.74
Lad	Brother	1.39	1.66	2.41
Crane	Implement	1.26	1.68	2.37
Journey	Car	1.05	1.16	1.55
Food	Rooster	1.18	0.89	1.09
Coast	Hill	1.24	0.87	1.26
Forest	Graveyard	0.41	0.84	1.00
Shore	Woodland	0.81	0.63	0.90
Monk	Slave	0.36	0.55	0.57
Coast	Forest	0.70	0.42	0.85
Lad	Wizard	0.61	0.42	0.99

Where,

$[0,1]$ determines the degree of influence of the generalizations.

(x) (y) are the reachable nodes shared by x and y

The shared nodes approach with similarity function discussed above complies with all the defined properties.

3.4.4 Weighted Shared Nodes Similarity

It is found that that when deriving similarity using the notion of shared nodes, not all nodes are equally important. If we want two concepts to be more similar when they have an immediate subsuming concept, we must differentiate and cannot simply define (c) as a crisp set.

(c) can be derived as follows. Let the triple (x, y, r) be the edge of type r from concept x to concept y; let E be the set of all edges in the ontology; and let T be the top concept, which means:

$$\alpha(T) = \{T\}$$

(c,c1,r)

A simple modification that generalizes (c) to a fuzzy set is obtained through a function $weight(r)$ that attaches a weight to each relation type r. we must differentiate and cannot simply define (c) as a crisp set.

The following is a generalization to fuzzy set based similarity [Andreasen *et al.*, 2005b][10], denoted as weighted shared node

$$\alpha(T) = \{1/T\}$$

(c,c1,r)

$$\cup \frac{c' \in \alpha(c1)}{c1} \alpha(c1) weight(r) \cdot \mu(c$$

Assigning weights to edges is very important, as it generalizes the measure so that it can be make use for different domains with different semantic relations. It also allows differentiating between the key ordering relation, ISA and the other semantic relations when calculating similarity. The weighted shared nodes measure complies with all the defined properties.

4. COMPARISON OF DIFFERENT SIMILARITY MEASURES

In this section we discuss about the results of comparison of the measures to human similarity judgments. The first human similarity judgment was done by Rubinstein and Goodenough [1965][11], using two groups totaling 51 subjects to perform synonymy judgments on 65 pairs of nouns and this in turn been the basis of the comparison of similarity measures. A subset of the noun pairs used in charles and miller experiments is shown on Table 1.

Miller and Charles [1991][12] repeated Rubinstein and Goodenough's original experiment, they used a subset of 30 noun pairs from the original list of 65 pairs, where ten pairs were from the high level of synonymy, ten from the middle level and ten from the low level. The correlation of these experiments was 0.97. The correlations for the resnik,Jiang and cornath,liu and St-Onge and leacock and chodorow are available in [18][19].

5. CONCLUSION

This paper has discussed the various approaches that could be used for finding similar concepts in an ontology and between ontologies.The purpose of this survey is to exploit the similarity methods for ontology based query expansion to aid better retrieval effectiveness of Information retrieval models. The experiments conducted by early researches provide better correlation values which gives promising direction of using them in Ontology based retrieval models. We are working for innovative similarity function which will combine the advantages of the similarity methods discussed in this paper and we will test it with ontologies of particular domain.

6.. REFERENCES

[1] Roy Rada, H. Mili, Ellen Bicknell, and M. Blettner. Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man, and Cybernetics, 19(1):17{30, January 1989.

[2] Michael Sussna. Word sense disambiguation for tree-text indexing using a massive semantic network. In Bharat Bhargava,

[3] Zhibiao Wu and Martha Palmer. Verbs semantics and lexical selection. In Proceedings of the 32nd annual meeting on Association for Computational Linguistics, pages 133{138, Morristown, NJ, USA, 1994. Association for Computational Linguistics.

[4] Claudia Leacock and Martin Chodorow. Combining local context and wordnet similarity for word sense identification. In Christiane Fellbaum, editor, WordNet: An Electronic Lexical Database. MIT Press, 1998.

[5] Philip Resnik. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language, 1999.

[6] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and K.J. Miller.[Resnik, 1995] Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. In IJCAI, pages 448{453, 1995.

[7] D. Conrath J. Jiang. Semantic similarity based on corpusstatistics and lexical taxonomy. In Proceedings on International Conference on Research in Computational Linguistics, Taiwan, pages 19{33, 1997.

[8] Dekang Lin. An information-theoretic definition of similarity. . Shavlik, editor, ICML, pages 296{304. Morgan Kaufmann, 1998. ISBN 1-55860-556-8.

[9] Graeme Hirst and David St-Onge. Lexical chains as epresentation of context for the detection and correction malapropisms. In Christiane Fellbaum, editor, WordNet: An Electronic Lexical Database. MIT Press, 1998.

[10] Troels Andreasen, Rasmus Knappe, and Henrik Bulskov. Domain-speci¯c similarity and retrieval. In Yingming Liu, Guoqing Chen, and Mingsheng Ying, editors, 11th International Fuzzy Systems Association World Congress, volume 1, pages 496{502, Beijing, China, 2005. IFSA 2005, Tsinghua University Press.

[11] Rubinstein and Goodenough, 1965] H. Rubinstein and J. B. a Goodenough. Contextual correlates of synonymy. Communications of the ACM, 8(10), 1965.

[12] George A. Miller and W. G. Charles. Contextual correlates of semantic similarity. Language and Cognitive Processes, 6(1):1{28, 1991.

[13] Dekang Lin. Using syntactic dependency as local context to resolve word sense ambiguity. In ACL, pages 64{71, 1997.

14] Roy Rada and Ellen Bicknell. Ranking documents with a thesaurus. JASIS, 40(5):304{310,1989.Timothy Finin, and Yelena Yesha, editors, Proceedings of the 2nd International Conference on Information and Knowledge Management, pages 67{74, New York, NY, USA, November 1993. ACM Press.

[15] Alexander Budanitsky, University of Toronto Graeme Hirst, University of Toronto Evaluating WordNet-based Measus Of Lexical Semantic Relatedness, Association for Computational Linguistics,2006

[16] Philip Resnik, Using Information Content to Evaluate Semantic Similarity in a Taxonomy, Sun Microsystems Laboratories, 1995

[17] Henrik Bulskov Styltsvig, Ontology-based Information Retrieval Computer Science Roskilde University, 1996

[18] A. Budanitsky. Lexical semantic relatedness and its application in natural language processing, 1999.

[19] A. Budanitsky. Semantic distance in wordnet: An experi- mental, application-oriented evaluation of ¯ve measures, 2001.