International Journal of Engineering Research and Development
***e**-ISSN: 2278-067X, **p**-ISSN: 2278-800X, www.ijerd.com*
Volume 3, Issue 3 (August 2012), PP. 40-44

Similarity Measure Based On Edge Counting Using Ontology

Vadivu Ganesan[1], Rajendran Swaminathan[2,] M.Thenmozhi[3]
[1]*Assistant Professor, SRM University, 603203, India*
[2]*Professor, SRM University, 603203, India*
[3]*Assistant Professor, SRM University, 603 203, India*

Abstract—Building the ontology from the scratch is a difficult process and there is no proper fully automated ontology construction methodology is available. But more and more ontologies are created and available in the web, reusing the existing ontology is reasonable for the ontology developers. Ontology reuse is one of the research issue which leads to ontology mapping, ontology merging and ontology integration. Ontology mapping process required to find the semantic similarity between the terms. Semantic similarity can be identified by calculating lexical similarity and conceptual similarity. Wu and Palmer developed a simple and good performance algorithm compared to other similarity measure measures. In this paper, a modified algorithm of Wu and Palmer is discussed and the results are compared with Wu and Palmer algorithm.

Keywords—Semantic web, ontology reuse, semantic similarity, Wu and Palmer algorithm.

I. INTRODUCTION

ONTOLOGY describes the concepts, their relationships and properties within their domain and it can be utilized both to offer automatic inferring and interoperability between applications. This is an appropriate vision for knowledge management. Ontology provides understandability of the structured information. With a common ontology, information that is spread out in many different applications and documents can be viewable in an easy way to understand and navigate. The ontology makes it possible to search both explicit and tacit knowledge, thereby bridges the gap between the tacit and explicit knowledge. The advantages of ontology are: knowledge sharing, logic inference and reuse of knowledge.

Ontology defines a common vocabulary for researchers who need to share information in a domain. It includes machine-interpretable definitions of basic concepts in the domain and relations among them.

In practical terms, developing ontology includes:

- Defining classes in the ontology,
- Arranging the classes in a taxonomic (subclass–superclass) hierarchy,
- Defining properties (or slots),
- Filling in the values for properties of instances.

Mapping is one of the sub processes of integration which is the process of building ontology in one subject and reusing it by one or more other subjects. The steps of mapping process are to identify the available ontologies, and then finding the possible terms to be mapped.

To find the terms to be mapped, semantic similarity between the ontology terms have to be calculated in automated way. Since for the large scale of data it is not possible to perform the manual mapping among the terms. Mapping of ontologies requires lexical mapping and conceptual mapping.

II. RELATED WORK

In [8], some of the integration challenges addressed in this paper were:

- Finding similarities and differences between ontologies in automatic and semi-automatic way
- Defining mappings between ontologies
- Developing an ontology-integration architecture
- Composing mappings across different ontologies
- Representing uncertainty and imprecision in mappings

Ontology-integration on the large scale will be possible only if there is significant progress in identifying mappings automatically or semi automatically.

In [5], types of semantic relatedness were discussed. Hierarchical taxonomy expressing IS-A (hypernymy / hyponymy) relation is considered to be the most appropriate for determining the semantic similarity. It can be used by two main approaches: edge-based and node-based.

A. Edge-Based Approach [5]

Edge-based approach measures minimal distance between concepts (synsets) in a hierarchical structure. Resnik[9] presents edge-based measure that converts the minimal length between concepts c1 and c2. It is given by:

$$sim^{e}_{R}(w1, w2) = 2 \times MAX - min_{c1,c2}len(c1, c2)$$

where MAX is maximum depth of the taxonomy and len(c1,c2) is length of the shortest path between concepts c1 and c2.

B. Node-Based Approach [5]

In addition to hierarchical taxonomy, node-based approach uses a large text corpus to compute probabilities of encountering an instance of concept c and then its information content.

Lin's similarity measure:

$$\mathrm{sim}^{n}_{L}(w_1, w_2) = \frac{2\log(p(lso(c_1, c_2)))}{\log(p(c_1)) + \log(p(c_2))},$$

where lso(c1, c2) is the lowest super-ordinate of word concepts c1 and c2.

In [2], the measure of Wu and Palmer [1] has the advantage of being simple to implement and have good performances compared to the other similarity measures. Nevertheless, the Wu and Palmer measure present the following disadvantage: in some situations, the similarity of two elements of an IS-A ontology contained in the neighborhood exceeds the similarity value of two elements contained in the same hierarchy.

In the field of the information retrieval which is largely based on the similarity identification measures between documents. The problem of those approaches is that they typically focus on the single words of a document ignoring the ontological relationships that exist between the words. We can distinguish three ways to determine the semantic similarity between objects in ontology. The first approach indicates the evaluation of the similarity by the information content (also called the node based approach). The second approach represents an evaluation of the similarity based on conceptual distance (also called edge based approach). The third approach is hybrid which combines the first two approaches.

Given an ontology Ω formed by a set of nodes and a root node (Root) (Fig. 1). Term1 and Term2 represent two ontology elements of which similarity to be calculated. The principle of similarity computation is based on the distance Depth1 and Depth2 from Root; Depth, which separates nodes Term1 and Term2 from the closest common ancestor Common Node.

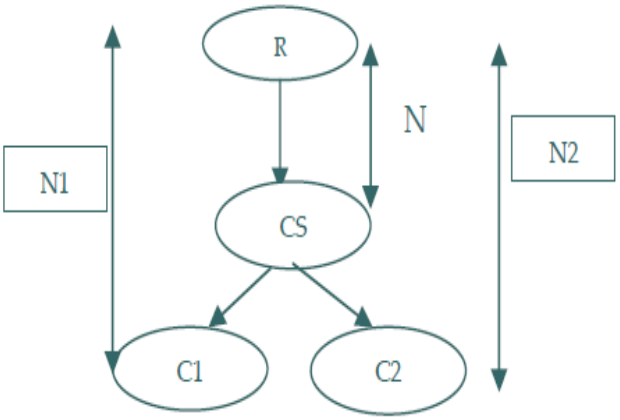

Fig.1 Example of Ontology Extract used in [1, 2]

The similarity measure of Wu and Palmer[1] is defined by the following expression:

$$SIM_{wp} = 2 * depth / \left(depth1 + depth2 \right);$$

The following is the code in Java to implement Wu Palmer:

```
int N1 = depthFinder.getShortestDepth(synset1 );
int N2 = depthFinder.getShortestDepth( synset2 );
double score = 0;
if (N1>0 && N2 >0) {
  score = (double)( 2 * N ) /
        (double)( N1 + N2);}
```

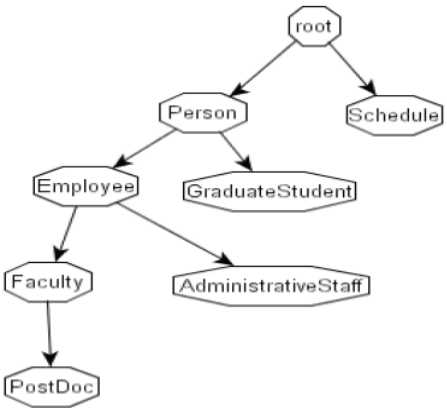

Fig. 2 An extract of UnivBench ontology used in [2]

Example 1: Let the ontology of Fig 2, C1, C2 and C3 the concepts "Person", "PostDoc" and "AdministrativeStaff". Sim_{wp} (C1, C2) =2*1/ (1+4) =0.4 and Sim_{wp} (C2, C3) =2*2/(4+3)=4/7=0.57.
The similarity values obtained by Wu and Palmer show that the neighbor concepts C2 and C3 are more similar than the concepts C1 and C2 located in the same hierarchy, which is problematic and inadequate within the semantic information retrieval. A new measure which is inspired from the advantages of Wu and Palmer work, whose expression is represented by the following formula [tbk]:

$$Sim_{tbk}(C1, C2) = \frac{2.N}{N1 + N2} * PF(C1, C2)$$

Let PF(C1, C2) be the penalization factor of two concepts C1 and C2 placed in the neighborhood.

$$PF(C1,C2) = (1-\lambda).(Min(N1,N2) - N) + \lambda.(|N1 - N2| + 1)^{-1}$$

Let N1 and N2 be the distances which separate nodes C1 and C2 from the root node, and N, the distance which separates the closest common ancestor of C1 and C2 from the root node. C1 and C2 are the concepts for which the similarity is computed. The coefficient λ is a Boolean value indicating 0 or 1, with 0 indicating two concepts in the same hierarchy and 1 indicating two concepts in neighborhood, respectively. Min(N1, N2) represent the minimum value between C1 and C2. To check the validity of the measure, the Fig.3 is used.

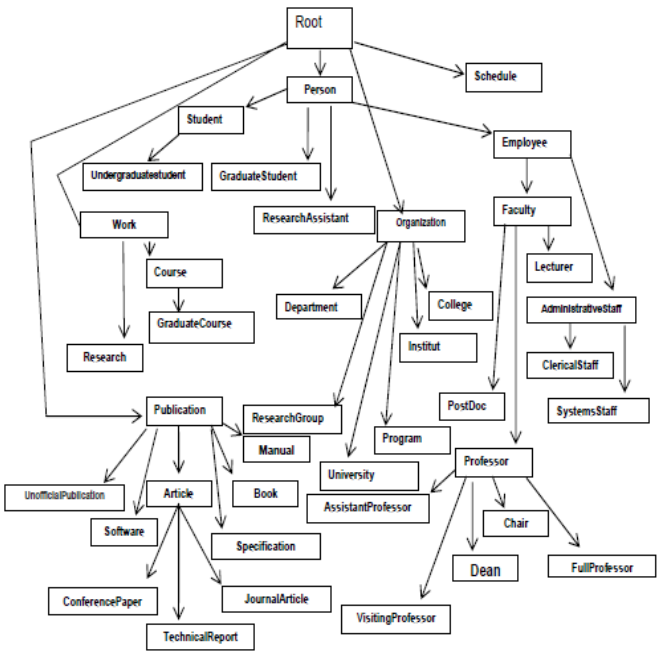

Fig. 3 Univ-Bench Ontology used in [2]

The Table I show the results of Wu and Palmer compared to tbk. And the conclusion is that the lower similarity value is given for close concepts compared to concepts in the same hierarchy based on the tbk algorithm.

Table I: Comparison of WuPalmer and tbk given in [2]

C1, C2	Sim_{wp}	Sim_{tbk}
Person, ResearchAssistant	0.66	0.66
VisitingProfessor, FullProfessor	0.8	0.8
VisitingProfessor, SystemsStaff	0.44	0.22
ResearchAssistant, Faculty	0.4	0.2
Chair, AdministrativeStaff	0.5	0.16
Research, GraduateCourse	0.4	0.2
SystemsStaff, Professor	0.5	0.5
SystemsStaff, Dean	0.44	0.22
Person, Schedule	0	0

III. PROPOSED METHDOLOGY

We have analyzed the edge based algorithms Wu and Palmer and tbk, the proposed algorithm will give better result. Both algorithms are using the distance up to Root node in calculating the similarity between the nodes.

In our proposed method, by considering the distance between the nodes to be compared are considered and not the distance from the Root node.

Fig.4 shows the variables used in the proposed algorithm. To calculate the similarity score, log function is used. The distance between the nodes to be compared is more than 10 is considered as not related terms and are not of interest.

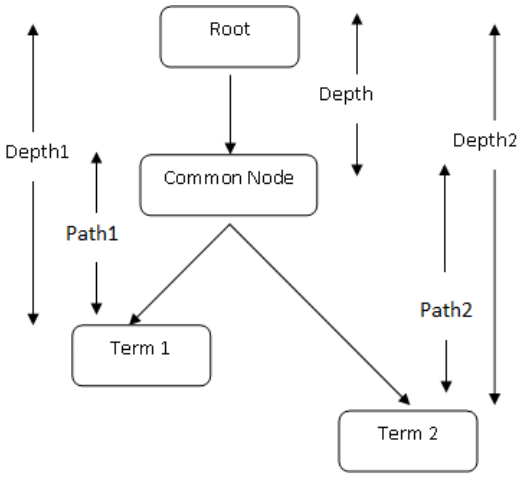

Fig. 4 concepts of proposed algorithm

```
int depth1 = depthFinder.getShortestDepth(synset1 );
int depth2 = depthFinder.getShortestDepth( synset2 );
double score = 0;
path1 = depth1 – depth;
path2 = depth2 – depth;
if (depth1>0 && depth2 >0) {
score = log(10-(path1+path2))
```

The above code is based on the proposed algorithm. Table II shows the comparison of Wu and Palmer, tbk and new algorithm: [based on the same Fig.3]

Table II : Wu and Palmer, tbk and new algorithm

Term1, Term2	**WuP**	**tbk**	**new**
Person, ResearchAssistant	0.66	0.66	0.950
VisitingProfessor, FullProfessor	0.80	0.80	0.903
VisitingProfessor, SystemStaff	0.44	0.22	0.698
ResearchAssistant, Faculty	0.40	0.20	0.840
Chair, AdministrativeStaff	0.50	0.16	0.770
Research, GraduateCourse	0.40	0.20	0.840
SystemStaff, Professor	0.50	0.50	0.770
SystemStaff, Dean	0.44	0.22	0.698

IV. RESULT ANALYSIS

Maximum similarity is 1 and minimum similarity is considered as 0 for analyzing the results.

Example 1: Person, ResearchAssistant

Person, ResearchAssistant are in the same hierarchy and the length difference is 1.

Example 2: VisitingProfessor, FullProfessor

These two terms are from a single parent and the distance length is 2, slightly lesser than the value given in example 1.

Example 3: VisitingProfessor, SystemStaff

The distance between these two terms are 5, therefore the similarity measure is so low compared to previous examples.

Example 4: ResearchAssistant, Faculty

The distance between these two terms are 3, therefore the similarity value is more than the value given in example 3. But both WuP and tbk are with larger value than example 3.

Example 5: Chair, AdministrativeStaff

The distance between these two terms are 4, therefore the similarity value is more than the value given in example 3 and less than example 4. But both WuP and tbk are giving different values.

Example 6: SystemStaff, Professor

The distance between these two terms are 4, therefore the similarity value is similar to example 5. But tbk is giving different value.

V. CONCLUSION

In this work we have presented an extension of similarity measure based on Wu and Palmer measure. We have compared our measure with Wu Palmer and tbk. The obtained results shows the relevance of the similarity measure between two concepts contained in a hierarchical ontology compared to the work of [1] and [2].

REFERENCES

[1]. Z. Wu and M. Palmer. "Verb semantics and lexical selection". In Proceedings of the 32nd Annual Meeting of the Associations for Computational Linguistics, pp 133-138. 1994.

[2]. T. Slimani, B. Ben Yaghlane, and K. Mellouli, "A New Similarity Measure based on Edge Counting" World Academy of Science, Engineering and Technology, PP 34-38. 2006.

[3]. Helena Sofia Pinto & Jo˜ao P. Martins, "Methodology for Ontology Integration", K-CAP'01, 2001, Victoria, British Columbia, Canada. ACM 1-58113-380-4/01/0010.

[4]. B. Chandrasekaran, J. Josepheson, V.R. Benjamins. Ontologies: What are they? Why do we need them? IEEE Intelligent Systems, 14(1):20-26, 1999.

[5]. Gabriela Polˇcicov´a and Pavol N´avrat, "Semantic Similarity in Content-Based Filtering", ADBIS 2002, LNCS 2435, pp. 80–85, 2002.
Springer-Verlag Berlin Heidelberg 2002.

[6]. Namyoun Choi, Il-Yeol Song, and Hyoil Han, "A Survey on Ontology Mapping", SIGMOD Record, Vol. 35, No. 3, Sep. 2006.

[7]. Octavian Udrea, Lise Getoor, Renée J. Miller, "Leveraging Data and Structure in Ontology Integration", SIGMOD'07, June 11–14, 2007, Beijing, China. Copyright 2007 ACM 978-1-59593-686-8/07/0006.

[8]. Natasha F. Noy, "What do we need for ontology integration on the Semantic Web Position statement", Proceedings of the Semantic Integration Workshop, Collocated with the Second International Semantic Web Conference (ISWC-03), 2003.

[9]. Philip Resnik, "Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language", Journal of Artificial Intelligence Research 11 (1999) 95-130.

[10]. Wordnet, http://wordnet.princeton.edu/