Journal of Theoretical and Applied Information Technology
20th January 2016. Vol.83. No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

CONCEPTUAL SIMILARITY MEASURE

[1]LABRIJI AMINE, [2]ABDELBAKI ISSAM, [3]REDDAHI NABIL, [4]ABDELOUHED NAMIR, [5]ABOUDOU ABDERRAOUF

Department of mathematics and informatic, Faculty of Sciences Ben M'SIK

University HassanII – Mohammadia, Casablanca, Maroc

E-mail: [1]labriji73@hotmail.com, [2]i.abdelbaki@gmail.com, [3]reddahi.nabil@gmail.com, [4]a.namir@yahoo.fr, [5]a.aboudou@gmail.com

ABSTRACT

The similarity is a problem that has been the subject of several research projects, particularly in the field of semantic information retrieval; this latter is based on ontologies for modeling knowledge, a similarity measure of ontology concepts is thus necessary in the main phases of information retrieval (Indexing, weighting, research, …).

We present through this article a similarity computing approach to arc based ontology of concepts. We assessed the similarity values obtained with those of the most used approaches, namely "The measure of Wu and Palmer" and "The measure of Rada and al". It shows that our measure is beneficial and provides a solution to the limits of existing approaches.

Keywords: *Ontologies, Semantic Web, Arcs Distances, Similarity Measure, Information Retrieval.*

1. INTRODUCTION

The issue of identifying the similarity in the ontologies and/or calculating semantic distances is considered as a widely used research subject by several areas such as consolidation, data mining, semantic web and especially in the information retrieval. The latter is based on measurements for identification of the similarity between documents [Bar, 1] [Sal, 2]. The problem with these approaches s is that they generally focus only on concepts ignoring the ontological relationships between them.

We can distinguish three ways to determine semantic similarity between objects in the ontology. The first approach measures the similarity by information content (also called the node based approach). The second approach represents an evaluation of the conceptual similarity based on the distance (also known as the edge based approach). The third approach is hybrid it combines the first two approaches. However the second approach is dependent on the ontology construction. In fact, in some situations we can obtain similarity value of two elements of an ontology contained in the neighborhood which exceeds the value of similarity of two concepts contained in the same hierarchy. This situation is inadequate within the information retrieval context. It is in this in this context that our work aims to propose a new similarity measure to overcome the problem mentioned above.

The objective behind the generation of a new similarity measure is to get realistic results for concepts not located in the same path. The rest of this article is organized on six sections. The first presents some fields of application of the similarity measure. In the second section, we will present the ontologies and a brief description of the ontology used. In the third and fourth sections, we present the principle of similarity measures and a discussion of some works identified in the literature. The fourth section presents a detailed presentation of our similarity measure with detailed examples. The experimental results of our prototype and a comparison with other works are presented in section 5. Finally, we conclude in the sixth section with some perspectives.

To demonstrate the importance of the similarity measure, we chose several application areas in order to clarify its use.

2. NATURAL LANGUAGE PROCESSING (NLP)

Several studies on the similarity measure were encouraged by the natural language processing (NLP). Among the works in this field we can cite: the work of [Pat, 3] uses the metric of semantic similarity for measuring semantic similarity between all the senses of a pair of words and

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

disambiguate them in a given context. [Mcc, 4] combined the use of a thesaurus automatically acquired from raw textual corpus and Wordnet (based on similarity metric) to find the predominant meaning of words in unstructured texts. The authors of the work [Gus, 5] applied semantic similarity measures of Wordnet to evaluate the relevance of expressions, given a specific dialogue, and automatically build summary of spoken dialogue. The work of [Hib, 6] studied the usefulness of semantic similarity in orthographic correction problem, where real orthographic errors are identified and corrected automatically.

2.1 Bioinformatics

A variation measure of similarity based on the information content is adopted to find a better way to organize and to query data from gene ontology (GO). The work of [Lor, 7] is interested in semantic similarity between proteins rather than the terms of the ontology GO, this is why he combined three similarity measures [Res, 8], [Lin, 9] and [Jiac, 10].

2.2 Web Services

Determining the similarity of semantic services provides useful information regarding their compatibilities. In the work of [Jef, 11], there is a proposal metrics to measure the similarity of semantic services annotated with an ontology OWL. The proposed similarity measure is due on intuition that similar objects share the most common descriptive information.

2.3 Links detection

In the work of [Che, 12], There is a description of the improvement of systems based on link detection history using the specific source of information and combining a couple of similarity measures. The adopted similarity measures of this work are represented by cosine, Hellinger, Tanimoto and clarity. Each of these measures catches different aspects of similarity of words in a document.

2.4 Ontology

Ontology means a set of semantic resources hierarchically linked. It describes the knowledge of a specific domain and presents relationships between concepts as well as giving the rules and axioms missing on semantic networks.

The purpose of ontology is to process semantically the information. In fact the semantic refers to a set of technologies which aim to make resources content understandable by the machine, through a formal metadata system, including the family of languages developed by the W3C (especially the semantic annotations, ontologies and thesauri).

2.5 User ontology:

There are several ontologies designed to list the content of web pages for easier navigation by users. We can cite online portals such as « Yahoo », « Mmagellan », « Lycos », and « ODP ». Seeing that ODP is the biggest and the most complete web directory edited by experts in the domain, it is used as a source of semantic knowledge in the process of information accessing. Our objective in this section is to represent each concept of ODP by a set of terms thus later serving in deriving the semantic representation of the user profile.

Figure 1: The Interface Of The Domain Ontology ODP

ODP's data are available in two files of type « RDF »: the first one contain the tree structure of the ontology ODP and the second list the resources or the associated web pages of each concept. The following figure shows an extract of the ODP architecture.

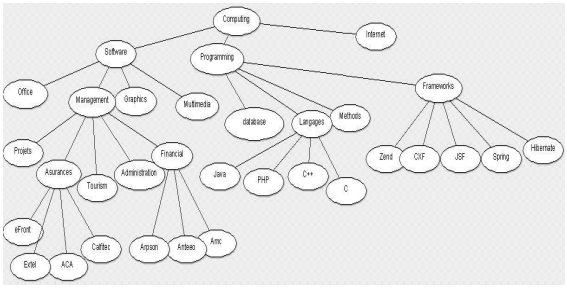


Figure 2: Extract Of The Architecture Of The Ontology ODP

- The concepts (also called categories): each concept of ODP represents an area of interest of web users and is manually associated by experts of domain to web pages, these latter are considered the most relevant to the concept. The

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

concepts are hierarchically organized where the top level concepts represent general concepts and the low level concepts represent specific concepts. Each concept of ODP is represented by a title and a description, generally, describing the content of associated web pages, and for each web page is also associated a title and a description describing its content.

- Links between concepts: The concepts of the ontology are linked with different relations types such as « is-a », « symbolic » and « related ». The links of type « is-a » allow to go hierarchically from generic to most specific concepts. The links of type « symbolic » support multi-classification of pages in several concepts and allow to the user to navigate between semantically linked concepts without recourse to general concepts. The links of type « related » are denominated by « see also » allow to point to concepts treating the same theme without sharing any common web pages.

3. SIMILARITY MEASURE

2.6 The similarity

The similarity notion in our context is not one that can be found in psychology or in mathematics. In social psychology, the similarity refers to how attitudes, values, interests and personality match between people. In mathematics, many equivalence relations (which are reflexive binary relations, symmetric and transitive) are called the similarity. For example these relations exist:

(i) In geometry: Two geometrical objects are similar if one is isometric with the result of uniform enlargement or contraction of the other. One can be obtained from the uniform enlargement or contraction of the other, with an eventual rotation (both have the same form), or more, applying a mirror effect (one in the same form as the mirror image of the other). For example, all circles are similar between them, all square are similar to each other, and all parables are similar to each other. On the other hand, both ellipsis and hyperboles are not similar to each other. Two triangles are similar if and only if they have the same 3 angles.

(ii) In linear algebra, two matrices A and B of size n × n are called similar if there is an invertible matrix P of the same size n × n satisfying $P^{-1}AP = B$.

(iii) In topology, the similarity is a function such as its value is greater when two points are closer (contrary to the distance, which is a measure of dissimilarity: closer are the points, smaller is the distance).

2.7 Semantic similarity

The semantic similarity is seen as that of topological similarity of mathematics, where it is associated with a function, called the similarity function. The definition of the similarity function may change depending on approaches, according to the desired properties. The value of this function is often between 0 and 1, allowing probabilistic opportunities interpretation of similarity. Possible common properties or characteristics of the function are positive characteristics, self-similar or maximum, symmetric and reflexive. We can also find other characteristics such as finitude or transitivity.

Definition 1. (Similarity)

The similarity

$$\begin{array}{rcl} S : O \times O & \rightarrow & [0,1] \\ (o,o) & \mapsto & S(o,o) \end{array}$$

is a function that associate to each pair of entities a real number expressing the similarity between these two entities such as :

1. $\forall a,b \in O,\ S(a,b) \geq 0$
2. $\forall a,b,c \in O, S(a,a) \geq S(a,b)$ et $S(a,a) = S(a,b) \Rightarrow a = b$
3. $\forall a,b \in O,\ S(a,b) = S(b,a)$
4. $\forall a,b,c \in O, S(a,b) = S(b,c) \Rightarrow S(a,b) = S(a,c)$

The dissimilarity is sometimes used instead of similarity. It is defined analogously to the similarity, except that is not transitive. We can distinguish three principle approaches for identifying similarity measures between taxonomy objects.

The first type is based on nodes [Res, 8], [Lin, 9] and [Jiac, 10]. It works under the banner of these approaches generally by using basically the information content to define the conceptual similarity. Furthermore, the similarity between the two concepts is obtained by the degree of sharing information.

The second type is solely based on hierarchy or edge distances [Lin, 9]. The problem with this

Journal of Theoretical and Applied Information Technology
20th January 2016. Vol.83. No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

approach is that the taxonomy arcs represent uniform distances that are all semantically linked with the same weight.

Finally, the hybrid approach [Jiac, 10], that combines the two approaches presented above. With these approaches, there are several ways to detect the conceptual similarity between two words in a hierarchical semantic network. In the next section we will present some measures contained in the second approach.

4. SIMILARITY MEASURES APPROACHES BASED ON ARCS

The most intuitive objects similarity measure in ontology is there distances [Rada and al, 13] [Lee, 14] [Wup, 15]. Obviously, an object X is more similar to an object Y than to an object Z. This similarity is evaluated by the distance separating the objects in the ontology. These measures use the hierarchical structure of the ontology to define the semantic similarity between concepts. Calculating distances in the ontology is based on a specialization objects graph. In each graph, the distance of the ontology must be characterized by the shortest path which involves a common ancestor or the smallest generalizer, potentially connecting two objects through common descendants. Among the classified works under this banner we can cite:

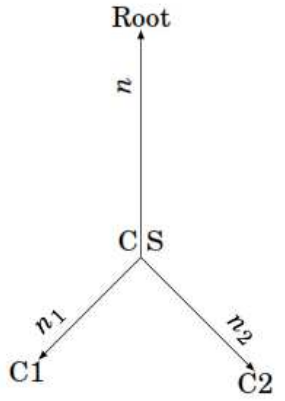
2.8 Rada and al measure

This measure [Rada, 13] is adopted in a semantic network and it is based on the fact that we can calculate the similarity based on hierarchical links « is-a ». To calculate the similarity between two concepts in the ontology, we must calculate the minimum number of arcs separating them. This measure, relative to metering edge between nodes by the shortest path, presents an average of the most obvious to evaluate the semantic similarity in a hierarchical ontology. Intuitively, this measure is based on the principle: an object A is judged more similar to an object B than an object C, if the distance from A to B within the graph is shorter than the distance from A to C. Rada and al. considers this distance, noted distedge(c1, c2), as the length of the shortest path between two concepts. The similarity between c1, c2 is defined by:

$$Sim_{Rad} = \frac{1}{dist_{edg}(C_1, C_2)}$$

2.9 Wu and Palmer measure

The principal of calculating similarity based on the count edge method is defined as follow; considering the ontology Ω formed by a set of nodes and a root node (R) (Fig. 1). C1 and C2 represent two elements of the ontology of which we will calculate the similarity. The principle of calculating similarity is based on the distance (n1 and n2) from nodes C1 and C2 to the closest common ancestor (CS) and the distance (n) from the closet common ancestor (CS) of C1 and C2 to the root node.

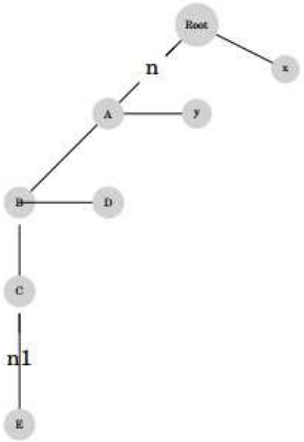


The similarity measure of Wu and Palmer is defined by the following expression:

$$Sim_{wp} = \frac{2N}{N1+N2+2N}$$

Proposal

Given a concept A, and B one of its children, C and D are two children of B, E is a descendant of the concept C



We have:

$$Sim_{WP}(E,A) < sim_{WP}E,D$$

Journal of Theoretical and Applied Information Technology
20th January 2016. Vol.83. No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Proof

Suppose:

$$Sim_{WP}(E,D) \leqslant sim_{WP}E,A$$

So:

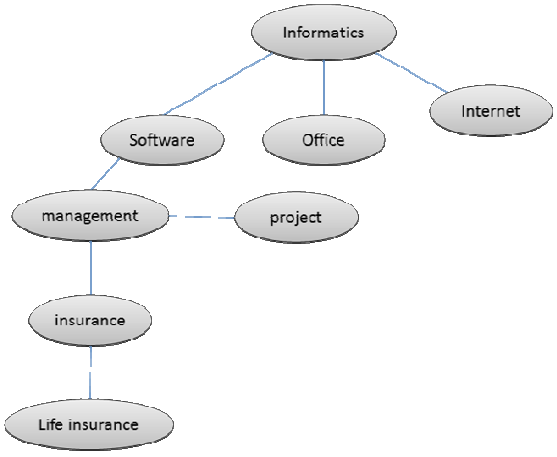
$$\begin{array}{rcl} \frac{2(n+1)}{n_1+2+2(n+1)} & \leqslant & \frac{2n}{n_1+2+2n} \\ (n+1)(n_1+2+2n) & \leqslant & n(n_1+2)+2n(n+1) \\ nn_1+4n+2n^2+n_1+2 & \leqslant & nn_1+4n+2n^2 \\ n_1+2 & \leqslant & 0 \end{array}$$

It's absurd, so we have:

$$Sim_{WP}(E,A) < sim_{WP}E,D$$

Example

Let the following ontology:



We denote by C1, C2 and C3 the concepts « Logiciel », « assurance vie » and « Projet ». Applying the Wu and Palmer measure, the similarity value is calculated as follows:

$$Sim_{wp}(C1,C2) = 2 * 1/(3+2) = 0.4$$
$$Sim_{wp}(C2,C3) = 2 * 2/(4+3) = 4/7 = 0.57$$

The similarity values obtained by Wu and Palmer measure show that the neighboring concepts C2 and C3 are more similar than the concepts C1 and C2 located on the same hierarchy which is inadequate in the context of semantic information retrieval.

The problem resulting from this measure is that the arcs in the ontology represent uniform distances (all the semantic links have the same weight). For this reason, we have adopted this measure as a basis for our work.

5. NEW SIMILARITY MEASURE

2.10 Ontology formalism

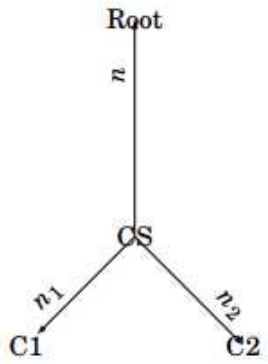
Let Ω an ontology which is an infinite set of classes and it is linked to a rooted tree. We designate by (C, P, HP, HC) the elements of Ω where C and P indicate, respectively, the set of classes and the set of properties contained in Ω. The hierarchies HP and HC indicate, respectively, the properties and class' hierarchy of Ω. The existing measure is interesting but has a limit because it aims essentially to detect the similarity between two concepts regarding to the distance of their smallest common subsumer (the nearest common concept). More general is the subsumption, less concepts are similar (and inversely). However, it did not gather the same similarity as the symbolic conceptual similarity (Consim). So, we can obtain Sim wp (A, D) given a descendant of A and B a sibling of A.

2.11 Formula of the new similarity measure

The measure of [Wup, 15] is interesting but is has a limit because it essentially aims to detect the similarity between two concepts to their distance to the smallest generalizer. This measure present the advantage of being fast in term of execution time, but it has a drawback of producing similarity value of two related concepts which exceed the value of two concepts in the same hierarchy. SO we can have with this measure

$$Simpw(A, f) < Simpw(A,B)$$

f is a child of A and B is a sibling of A, which is inadequate according to our sense in the context of information retrieval where we have to bring all the children of a concept (i.e. query) before its neighbors.



Let n1 and n2 the distances that separate the distinct nodes C1 and C2 from their closest common ancestor and n, the distance that separate the closest common ancestor of C1 and C2 from the

Journal of Theoretical and Applied Information Technology
20th January 2016. Vol.83. No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

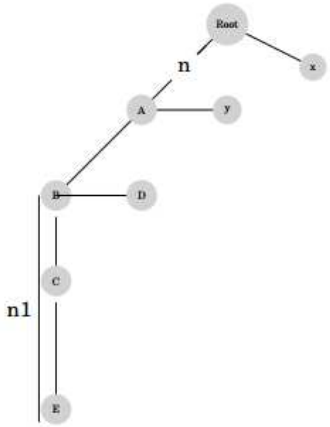
root node. We propose a new measure inspired from the advantages of the works of [Wup, 15], whose expression is represented by the following formula:

$$Sim_{lab}(C1,C2) = \frac{2n}{n1+n2+2n+n.n1.n2}$$

This new measure solves the problem of the Wu and Palmer measure.

Proposal

Given a concept A, and B one of its children, C and D are two children of D and E a descendant of the concept C



Proof

We suppose:

$$Sim_{lab}(E,A) < sim_{lab}E,D$$

So:

$$\begin{aligned} \frac{2n}{n_1+1+2n} &< \frac{2(n+1)}{n_1+1+2(n+1)+(n+1)n_1} \\ n(n_1+1+2(n+1)+(n+1)n1) &< (n+1)(n_1+1+2n) \\ n^2n_1+nn_1 &< n_1+1 \\ (n^2+n-1)n_1 &< 1 \end{aligned}$$

It's absurd, because n ≥ 1 and n1 ≥ 2, so we have:

$$Sim_{lab}(E,D) \leqslant sim_{lab}(E,A)$$

2.12 Property of the similarity measure proposed

In this section, we list some properties of the similarity measure. These properties depend on particular application, sometimes a property will be useful, while other desirable. The similarity function we propose ensures the following properties:

Given three concepts A, B and C of the ontology:

1. $Sim_{lab}(A,B) \geqslant 0,$
2. $Sim_{lab}(A,A) = 1,$
3. $Sim_{lab}(A,B) = Sim_{lab}(B,A),$
4. $Sim_{lab}(A,B) = 1 \Rightarrow A = B,$
5. $Sim_{lab}(A,B) + Sim_{lab}(A,C) \geqslant Sim_{lab}(B,C),$
6. $Sim_{lab}(A,B) + Sim_{lab}(B,C) \geqslant Sim_{lab}(A,C).$

2.13 Relevance of the similarity measure

In our context, a similarity measure is relevant, if it has a value for each couple of concepts (A, Bi) contained in the same hierarchy, which is always greater than or equal to these same concepts and related concepts (A, Ci). That is for every Bi descendant of A and for all related concepts Ci of A, we have

$$Sim_{lab}(A,B_i) \geqslant Sim_{lab}(A,C_i).$$

6. EXPERIMENTAL RESULTS

The objective of this work is to implement and to test a method of generating a new similarity measure that can advance researches in the domain of ontologies and simulation of conceptual distances.

We have tested our measure with an extract of the pedagogical domain ontology entitled univbench[1]. The following figure, presents an extract of this ontology:

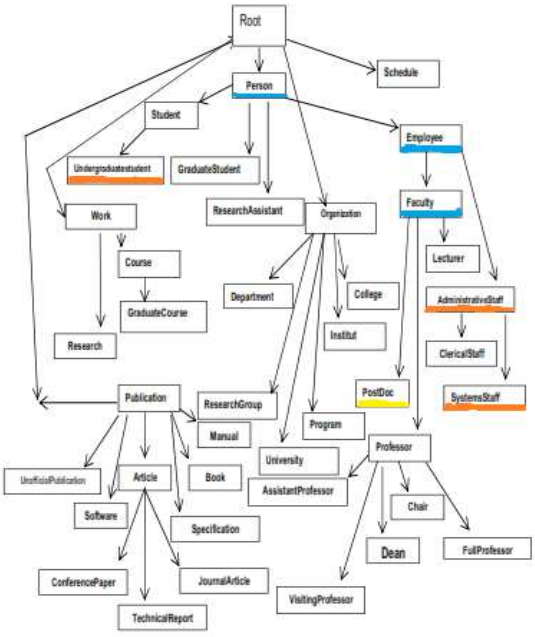


Figure 3: Extract Of The Ontology Univbench

1 : http://www.lehigh.edu/~zp2/2004/0401/univ-bench.owl

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

We calculated similarities of the concept « Post-Doc » (in yellow) with concepts of the same hierarchy (in blue) and with others in the neighborship (orange). The following table presents the results of similarity according to our measure and the measures of « Wu Palmer » and « Rada and al »:

Postdoc	Data	Wu and Palmer	New measure	Rada and al
Faculty	n = 3, n1 = 1 n2 = 0	0,85	0,85	0,5
Employer	n = 2, n1 = 2 n2 = 0	0,66	0,66	0,33
Person	n = 1, n1 = 3 n2 = 0	0,4	0,4	0,5
Administrative staff	n = 2, n1 = 2 n2 = 1	0,57	0,36	0,33
System staff	n = 2, n1 = 2 n2 = 1	0,5	0,25	0,5
Student	n = 1, n1 = 1 n2 = 3	0,33	0,22	0,5
Undergraduate student	n = 1, n1 = 2 n2 = 3	0,28	0,15	1

Table: Several Similarity Measures Results

The similarity values of our measure coincide with that of « Wu and Palmer » in the concepts of same hierarchy (Faculty, Employer, Person). However, our measure leads to smaller similarity values for concepts in the same neighborhood (Administrative staff, System staff, Student, Undergraduate student).

Seeing that the distance between subsumes concepts increase, we obtain smallest similarity values. A comparison of the relevance of our measure with that of Wu and Palmer is represented in the following figure.

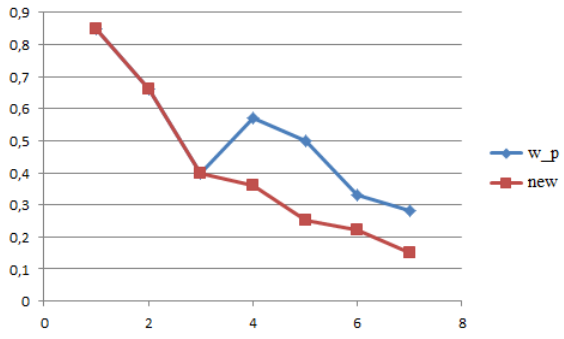


Figure 4: Comparison Between The New Measure And That Of Wu And Palmer

7. CONCLUSION

In this work we presented a new similarity measure based on arcs. We compared it with the « Wu and Palmer » measure and also the « Rada and al » both considered as the most used.

Experimental results show that the new measure improves the similarity between concepts in the same neighbors. The measure we defined can be used in several domains, in a first time we plan to use it to measure similarity between users' profiles.

REFERENCES

[1] R.Baeza-Yates et B.Ribeiro-Neto. Modern Information Retrieval. ACM Press; Addison-Wesley: New York; Harlow, England; Reading, Mass., 1999.

[2] G.Salton et M. J.McGill, Introduction to modern information retrieval. McGraw-Hill. New York, 1983.

[3] P. Siddharth, S. Banerjee et T. Pedersen. Using measures of semantic relatedness for word sense disambiguation. In Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City, Mexico, pp. 241-257. 2003.

[4] M. Diana, R. Koeling, J. Weeds et J. Carroll. Finding predominant senses in untagged text. In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, Barcelona, Spain, pp. 280 – 287, 2004.

[5] I. Gurevych et M. Strube. Semantic similarity applied to spoken dialogue summarization. In Proceedings of the 20th International Conference on Computational Linguistics, Geneva, Switzerland, 23 - 27, pp. 764-770, 2004.

[6] G. Hirst et A. Budanitsky Correcting real-word spelling errors by restoring lexical cohesion. Natural Language Engineering 2004.

[7] P.W. Lord, R.D. Stevens, A. Brass et C.A.Goble. Semantic Similarity Measures as Tools for Exploring the Gene Ontology. Pacific Symposium on Biocomputing 8, pp.601-612, 2003.

[8] P. Resnik. Using information content to evaluate semantic similarity in taxonomy. In Proceedings of 14th International Joint Conference on Artificial Intelligence, Montreal, 1995.

[9] D. Lin. An Information-Theoretic Definition of similarity. In Proceedings of the Fifteenth International Conference on Machine Learning

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

(ICML'98). MorganKaufmann: Madison, WI, 1998.

[10] J. Jiang et D. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In Proceedings of International Conference on Research in Computational Linguistics, Taiwan, 1997.

[11] H. Jeffrey, L. William et D. John. A Semantic Similarity Measure for Semantic Web Services. In proceedings of WSS05. 2005.

[12] F. Chen, A. Farahat et T. Brants. Multiple Similarity Measures and Source-Pair Information in Story Link Detection. In Proceedings of HLT, 2004.

[13] R. Rada, H. Mili, E. Bichnell et M. Blettner, Development and application of a metric on semantic nets. IEEE Transaction on Systems, Man, and Cybernetics: pp 17-30. 1989.

[14] J.H.Lee, M.H.Kim et Y.J.Lee. Information Retrieval Based on Conceptual Distance in IS-A Hierarchy. Journal of Documentation 49, pp. 188-207, 1993.

[15] 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.