World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:2, No:11, 2008

A New Similarity Measure based on Edge Counting

T. Slimani, B. Ben Yaghlane, and K. Mellouli

***Abstract*—**In the field of concepts, 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 [2]. 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. This situation is inadequate within the information retrieval framework. To overcome this problem, we propose a new similarity measure based on the Wu and Palmer measure. Our objective is to obtain realistic results for concepts not located in the same way. The obtained results show that compared to the Wu and Palmer approach, our measure presents a profit in terms of relevance and execution time.

***Keywords*—**Hierarchy, IS-A ontology, Semantic Web, Similarity Measure.

I. Introduction

THE question of similarity identification and/or the computation of semantic distances are regarded as a research subject highly investigated in the fields of data processing, Artificial Intelligence, and linguistics. In particular, the field of the information retrieval which is largely based on the similarity identification measures between documents [3][4]. 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. The problem of the second approach is dependent on the ontology construction. Furthermore, this approach, adopting IS-A ontology, presents the following disadvantage: in some situations, we can obtain a similarity value of two elements of an ontology contained in the neighbourhood which exceeds the value of similarity of two concepts contained in the same hierarchy. This situation is inadequate within the information retrieval framework.

Manuscript received September 23, 2006.

Authors are with IHEC Carthage, Carthage Presidency 2016, Tunisia (thabet.slimani@issatm.rnu.tn, boutheina.yaghlane@ihec.rnu.tn, khaled.mellouli@ihec.rnu.tn).

In order to overcome this problem, we propose, in this paper, a new similarity measure giving realistic results and closer relations to reality for concepts not located in the same path. The paper presents a similarity measure that is suited for comparing concepts in ontology. Although finding similar concepts is a core task in the area of ontology alignment/merging [5][6]. The proposed measure can be adopted effectively in this field.

The remainder of this paper is organized as follows: Section 2 reviews the literature on similarity measures. Section 3 simulates similarity measure to the conceptual proximity. Section 4 is a detailed presentation of our similarity measure with some examples. The experimental results of our prototype and a comparison with other works are presented in section 5. Finally, section 6 concludes with some future perspectives.

II. State of the Art

We can distinguish three main approaches for the similarity identification measures between the taxonomy objects. The first type is based on the nodes [2] [7] [8]. Works under the banner of these approaches used the typically information-based content to determine the conceptual similarity. Moreover, the similarity between two concepts is obtained by the degree of sharing information.

The second type is based only on the hierarchy or the edge distances [1][9][10][11]. The problem with this approach is that the taxonomy arcs represent uniform distances, i.e. all the semantic links have the same weight. Finally, the hybrid approach [12][13][14][15] which combines the two approaches presented above. With these approaches, there exist several manners of detecting conceptual similarity of two words in a hierarchical semantic network. The following section presents some measures which are listed under the second approach.

A. Wu and Palmer Measure

The principle of similarity computation is based on the edge counting method which is defined as follows:

Given an ontology Ω formed by a set of nodes and a root node (R) (Fig. 1). C1 and C2 represent two ontology elements of which we will calculate the similarity. The principle of similarity computation is based on the distance (N1 and N2) which separates nodes C1 and C2 from the root node and the distance (N) which separates the closest common ancestor

(CS) of C1 and C2 from the node R. The similarity measure of Wu and Palmer is defined by the following expression:

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

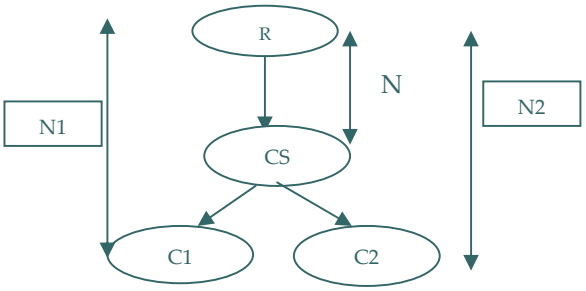

Fig. 1 Example of Ontology Extract

The problem resulting from this measure is that the arcs in ontology represent uniform distances (i.e. all the semantic links have the same weight). A comparison between the methods of similarity measure is carried out by [2]. This comparison reveals that the Wu and Palmer measure [1] has the advantage of being simple to calculate, in addition to the performances which it presents while remaining as expressive as the others. For this reason we have adopted this measure as a base for our work.

B. Rada et al. Measure

This measure [10] is adopted in a semantic network and it is founded on the fashion that we can calculate the similarity based on the hierarchical links “IS-A”.

To calculate the similarity of two concepts in ontology, we must calculate the number of the minimal arcs which separate them. This measure, based on the edge counting between nodes by the shortest way, presents a mean of the most obvious to evaluate the semantic similarity in a hierarchical ontology.

C. Ehrig et al. Measure

A work of similarity measure based on ontology was introduced by [11]. This work presents three layers: data, ontology and context. The similarity of the entities is measured on the data level by considering the data values of simple or complex types (integer, strings). The semantic relationships between the entities are measured on the level of ontology layer. The context layer specifies how the ontology entities are used in a certain external context, more specifically, the application context. All the previously listed similarities are calculated as function amalgams which combine the similarity measure of the individual layers.

III. Similarity Measure and Simulation of Conceptual Similarity

In Fig. 2, we present a graph representing a hierarchy of the concept. This graph represents an ontology extract of pedagogic field.

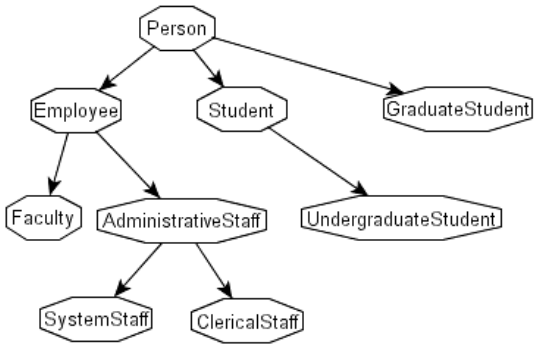

Fig. 2 Graph of concepts hierarchy

The concepts contained in this ontology represent intuitively a set of the varied conceptual distances if they are compared between them.

As an example the two concepts "Student" and "GraduateStudent" presents similarity value equal to 0 in the case of the use of traditional similarity measure, which includes external information with the hierarchy such as measure of [15][16]. On the other hand, the adoption of an approach based on the hierarchy gives a similarity measure different from 0 for these two same concepts. Moreover, the similarity value of two concepts "Student" and "GraduateStudent" are lower than the concepts "Student" and "UndergraduateStudent". However, we judge that the concept "Student" is closer with the concept "UndergraduateStudent" than the concept "Graduate Student".

These precise details are very interesting for the research of the semantic similarities of concepts set contained in ontology. These intuitive distances can be used, for example, to the improvement of the search engines on the level of the effectiveness and precision of answers to the user’s requests. The simplest structure supporting the reasoning on the hierarchy of types is that which can be found in a support of conceptual graphs.

In this structure, the IS-A links group the types according to the definitional characteristics which they share. The arrows presented in Fig. 3 present the relation IS-A from a superclass to its subclasses.

IV. Formula and Model of Proposed Similarity Measure

A. Ontology Formalism

Let Ω be an ontology which is a finite set of classes and seems to be **equated with a rooted tree**. We denote by (C, P, H^P, H^C) the elements of Ω where C and P indicate, respectively, the set of classes and the set of properties contained in Ω. The hierarchies H^P and H^C indicate, respectively, the hierarchy of properties and the hierarchy of classes of Ω.

The measure of [1] is interesting but presents a limit because it primarily aims to detect the similarity between two concepts compared to the distance of their least common

subsumer. The more this subsuming is general, the less similar they are (and conversely). However, it does not collect the same similarity as the symbolic conceptual similarity (conSim). Thus we can obtain Sim_{wp} (A, D) < conSim (A, B), D being one descendant of A and B one of the brothers of A. This situation is inadequate within the information retrieval framework where it is necessary to turn up all descendants of a concept (i.e request) before its vicinity.

For example, we can obtain with this measure, a value of similarity between the concept "PostDoc" and "AdministratifStaff" which exceeds the value of similarity between "Person" and "PostDoc". However, this measure offers a higher similarity between a concept and its vicinity compared to this same concept and a concept contained in the same path (see example 1).

Example 1: Let the ontology of figure 3, we indicate by 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.

B. Measure Formula

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. We put forward a new measure which is inspired from the advantages of [1] work, whose expression is represented by the following formula:

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

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.

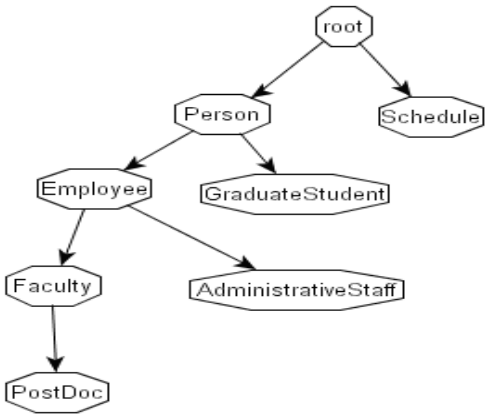

Fig. 3 An extract of UnivBench ontology

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.

The ratio *PF(C1, C2)* =1 if C1 is ancestor of C2 or the reverse. However, with this formula, we will penalize only the nodes that are in a neighborhood.

Our formula ensures a similarity Sim_{tbk}(C1, CV) always lower or equal to Sim_{tbk}(C1, CF), such as CV is a nearby node (not included in the same hierarchy of C1) and CF is a node placed in the same hierarchy of C1.
In the formula of *PF*(C1, C2) we have added the 1 outside the absolute value for the distance between C1 and C2, i.e., 1 / (|N1 – N2| + 1), because otherwise there could be a division by 0 in case N1 = N2.

C. Property of Proposed Similarity Measure

In this section we enumerate some properties of similarity measure [17]. These properties depend on a particular application; sometimes a property will be useful, sometimes it will be undesirable. The function of similarity which we propose ensures the following properties: Being given three concepts A, B and C of ontology:

1) ***Nonnegativity***: Sim_{tbk} (A, B)≥0,
2) ***Identity***: Sim_{tbk} (A, A) = Sim_{tbk}(B, B) =1;
3) ***Symmetry***: Sim_{tbk}(A, B) = Sim_{tbk}(B, A);
4) ***Uniqueness***: Sim_{tbk} (A, B) =1 implies A=B;
5) ***Strong triangle inequality***:
   Sim_{tbk}(A, B) + Sim_{tbk}(A, C)≥ Sim_{tbk}(B, C),
6) ***Triangle inequality***:
   Sim_{tbk}(A, B) + Sim_{tbk}(B, C)≥ Sim_{tbk}(A, C).

D. Relevance of Similarity Measure

In our context, a similarity measure is relevant, if it presents a value for each couple of concepts (A, Bi) contained in the same hierarchy, which is always higher or equal to this same concepts and any neighboring concept (A, Ci). i.e. ∀ concept Bi descendant of A and ∀ concept Ci neighbors of A, there exist Sim_{tbk} (A, Bi) ≥ Sim_{tbk} (A, Ci).

V. Experimental Results

The purpose of this work is to implement and analyze a generation process of a new similarity measure which can advance research in the ontology field and the simulation of conceptual distances. With this intention, it was necessary to develop a prototype to evaluate our work. The ontology, on which such calculations were made, is the ontology of pedagogic field which is entitled univ-bench[1]. This ontology is used to describe data concerning universities and their

[1] Accessible under LUBM benchmark with http://www.lehigh.edu/~zp2/2004/0401/univ-bench.owl

departments. The choice of this ontology is justified by the fact that it presents a field of which users can be familiar. This ontology was developed with the OWL language [18] and whose development is made for benchmarking reasons. This ontology contains 43 classes and 32 properties (including 25 properties of objects and 7 properties of the data type) (see Fig. 4).

In OWL ontology, each object is described by definite reports/links RDF [19]. Let O an object in an ontology OWL. O is characterized by a description set which contains all reports/links which describes to him. A description set for *O* is defined by: *Descr(O) = {(S, p, O)}*. Let (S, p, O) be an RDF triplet, formed by the predicate p, the object O and the subject S in *O*. Descr(O) contains all RDF reports/links in *O*.

RDF (Resource Framework Description) is used today as a standard for the metadata exchange between various applications. It makes it possible to facilitate the work of the search engines to find the efficient documents. RDF class (rdfs:Class) is identified by an URI. For example, "Professor" is an URI identifying class resources: Professors. There exist a triplet [#Professor, rdf:type, rdfs:Class].

RDFS language is used in our application to model the concepts contained in UnivBench ontology. RDF Classes and their interrelationships are obtained by RDQL language. RDQL [20] is the RDF interrogation language in the Jena models [21], with the joint similarity [22] which does not only provide the precise results of a request but also of the similar results.

RDQL is an implementation of the SQL query language for RDF. It treats an RDF model as data and provides a request with a model of triplet and offers the possibility of the constraints at the level of a simple RDF model.

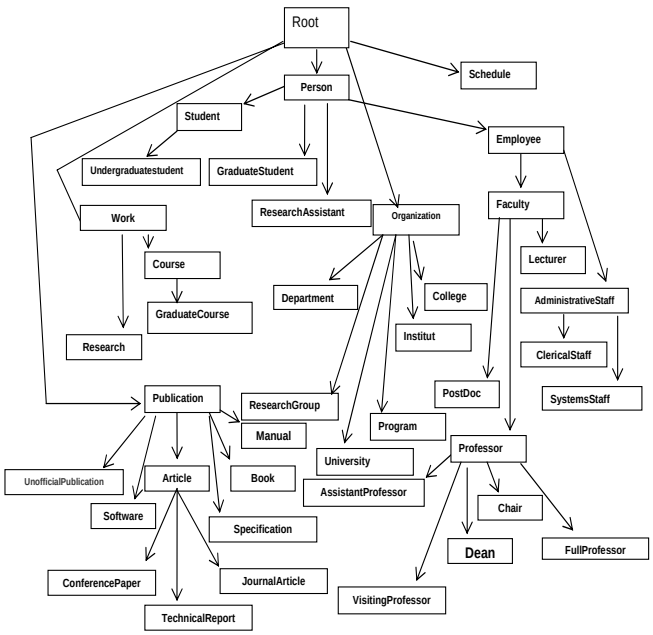

Fig. 4 Univ-Bench Ontology

To check the validity of our measure, it is judicious to test its speed of calculation compared to the measure of Wu and Palmer which was considered to be the fastest in terms of similarity generation time. The impact of the modification of the Wu and Palmer measure and the result with our measure must be evaluated to judge its relevance.

The developed prototype was based on PHP language and JavaScript. The ontology creation with RDF-Schema is based on Rap-rdf[2] API for PHP. RAP is a software package for the analysis (*parsing*), research, handling and serializing the RDF model.

In Table I, we have chosen pairs of concepts contained in ontology, in order to calculate the similarity values, respectively with our measure (Sim_{tbk}) and the measure of Wu and Palmer (Sim_{wp})[1].

TABLE I
EXPERIMENTAL RESULTS COMPARING OUR MEASURE (SIM_{TBK}) TO THE WU AND PALMER MEASURE (SIM_{WP})

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

Our measure is advantageous because it leads to a lower similarity value for close concepts compared to concepts in the same hierarchy.

The relevance of our measure compared to the Wu and Palmer measure is localized at the level of two concepts located in a hierarchy from which the subsuming concept [3] is different. As the distance between the direct subsuming concepts increases, lower similarity values are obtained. A comparison of the relevance of our measure compared to the Wu and Palmer measure is represented by Fig. 5. The obtained results show that there is an increase in the relevance brought by our measure.

In this work, we have treated only the case of two concepts. Certainly, it is possible to calculate the similarity between a set of concepts. This can be treated in a forthcoming work.

[2] http://www.wiwiss.fu-berlin.de/suhl/bizer/rdfapi/index.html
[3] A concept C1 is subsumed by C2 if C2 is the father and C1 the son.

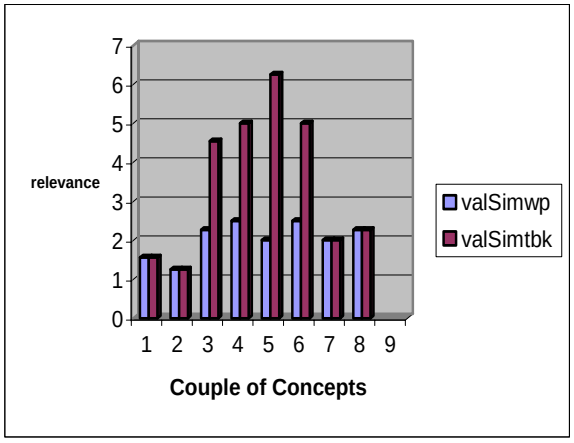

Fig. 5 Comparative histogram of the effectiveness of our measure compared to the Wu and Palmer measure

VI. Conclusion

In this work we have presented an extension of similarity measure based on Wu and Palmer measure. We have compared our measure with a measure which was regarded as the fastest computing. The obtained results show several advantages: The presented measure makes possible to increase the relevance of the similarity measure between two concepts contained in a hierarchical ontology compared to the work of [1]. The experimental results clearly show that the produced measure ensures at the same time the computing speed and the relevance of the produced values for the similarity between two concepts.

The similarity measure which we defined is advantageous since it provides realistic and adequate similarity values for all the ontology objects. The relevance of this measure increases, moreover, in the case of a hierarchical ontology, what makes it possible to give a clearer precision for the relations. This can be adopted in the field of the identification of semantic associations where the current approaches related to associations do not give a precision association accuracy degree. The utility of their use is to quantify associations (in the interval [0, 1]). For example, when an association has as value 0.8, it means that there is a confidence of 80% that the objects which form association are in direct or indirect relation.

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] D. Lin. "An Information-Theoretic Definition of similarity". *In Proceedings of the fifteenth International Conference on Machine Learning* (ICML'98). Morgan-Kaufmann: Madison, WI, pp.296-304. 1998.

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

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

[5] N.F. Noy and M. Musen. "PROMPT: Algorithm and Tool for Automated Ontology Merging and Alignment". In Proceedings of AAAI-2000, Austin, Texas. MIT Press/AAAI Press, 2000.

[6] M. Ehrig, S. Staab, Y. Sure. "Bootstrapping Ontology Alignment Methods with APFEL". International Semantic Web Conference 2005. pp. 186-200.

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

[8] N. Ho and F. Cédrick. "Lexical Similarity based on Quantity of Information Exchanged-Synonym Extraction". *In the Proceeding of Conf. RIVF'04*, February 2-5, 2004. Hanoi, Vietnam.

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

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

[11] M.Ehrig, P.Haase, M.Hefke, and N.Stojanovic. "Similarity for ontology-a comprehensive framework". In *Workshop Enterprise Modelling and Ontology*: *Ingredients for Interoperability*, 2004.

[12] 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.

[13] C. Leacock and M. Chodorow. "Combining Local Context and WordNet Similarity for Word Sense Identification. In *WordNet":* An Electronic Lexical Database, C. Fellbaum, MIT Press, 1998.

[14] P. 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. pp. 95-130. 1999.

[15] T. Eiter, and H. Mannila. "Distance measures for point sets and their computation". In *Acta Informatica Journal*, 34, 1997.

[16] J.Green, N.Horne, E.Orlowska and P. Siemens. "A Rough Set Model of Information Retrieval". *Theoretica Infomaticae* 28, pp 273-296, 1996.

[17] R. C. Veltkamp, and L.J. Latecki. "Properties and Performances of Shape Similarity Measures". 2006.

[18] M. Dean and G. Schreiber ed. "OWL Web Ontology Language Reference. W3C *Recommendation". 10 February 2004*. http://www.w3.org/TR/2004/REC-owl-ref-20040210/.

[19] G.Klyne and J.Carroll. "Web services description language (wsdl)1.1". http://www.w3.org/TR/rdf-concepts/, 2004.

[20] A.Seaborne. "RDQL - A Query Language for RDF", *W3C Member Submission*, 9 January 2004. http://www.w3.org/Submission/RDQL/.

[21] B.McBride. "Jena: Implementing the RDF Model and Syntax Specification". In *Proceedings of the Second International Workshop on the Semantic Web. SemWeb'2001.* May 2001.

[22] W. W. Cohen. "Data Integration Using Similarity Joins and a Word-Based Information Representation Language". *ACM Transactions on Information Systems,* Vol. 18, No. 3, July 2000.