International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue III, March 2018- Available at www.ijraset.com

A Hybrid Approach for Computing Semantic Similarity of Concepts in Knowledge Graphs

Harshal Wanjari[1], Prof. Nutan Dhande[2]
[1, 2] *Department of CSE, ACE Nagthana Wardha MH India*

Abstract: ***This paper presents a method for measuring the semantic similarity between concepts in Knowledge Graphs (KGs) such as WordNet and DBpedia. Previous work on semantic similarity methods have focused on either the structure of the semantic network between concepts (e.g. path length and depth), or only on the Information Content (IC) of concepts. We propose a semantic similarity method, namely wpath, to combine these two approaches, using IC to weight the shortest path length between concepts. Conventional corpus-based IC is computed from the distributions of concepts over textual corpus, which is required to prepare a domain corpus containing annotated concepts and has high computational cost. As instances are already extracted from textual corpus and annotated by concepts in KGs, graph-based IC is proposed to compute IC based on the distributions of concepts over instances. Through experiments performed on well known word similarity datasets, we show that the wpath semantic similarity method has produced statistically significant improvement over other semantic similarity methods. Moreover, in a real category classification evaluation, the wpath method has shown the best performance in terms of accuracy and F score.***

I. INTRODUCTION

Semantic similarity is a metric defined over a set of documents or terms, where the idea of distance between them is based on the likeness of their meaning or semantic content as opposed to similarity which can be estimated regarding their syntactical representation (e.g. their string format). These are mathematical tools used to estimate the strength of the semantic relationship between units of language, concepts or instances, through a numerical description obtained according to the comparison of information supporting their meaning or describing their nature.The term semantic similarity is often confused with semantic relatedness. Semantic relatedness includes any relation between two terms, while semantic similarity only includes "is a" relations. For example, "car" is similar to "bus",but is also related to "road" and "driving".

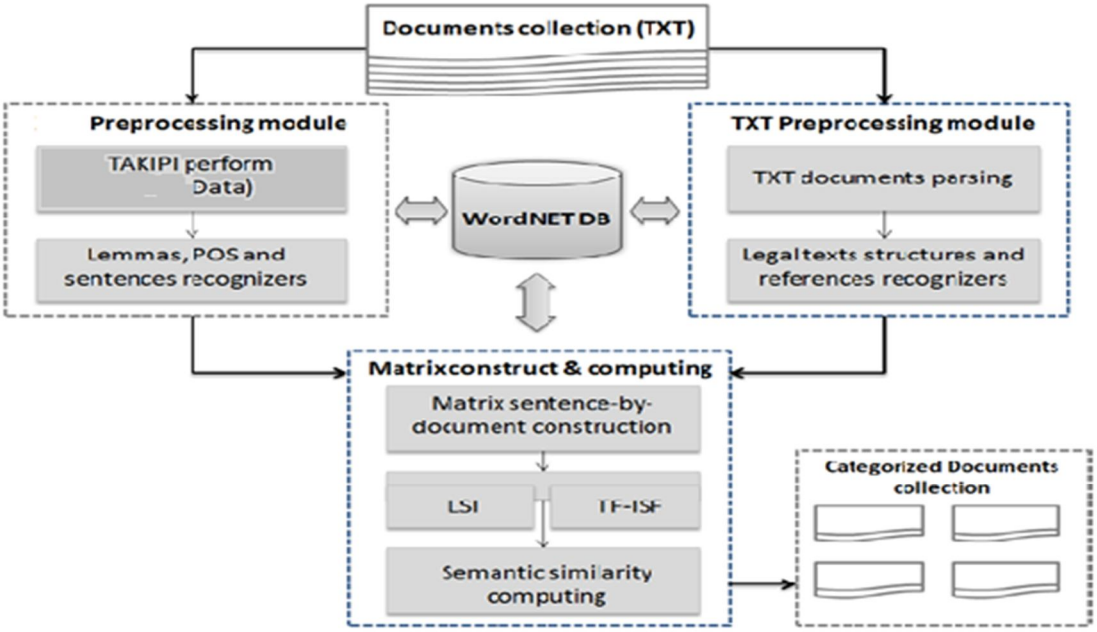

Figure.1.1 Architecture of Computing Semantics Similarity

To proposed a method for measuring the semantic similarity between concepts in Knowledge Graphs (KGs) such as WordNet and DBpedia. Previous work on semantic similarity methods have focused on either the structure of the semantic network between concepts (e.g. path length and depth), or only on the Information Content (IC) of concepts. We propose a semantic similarity method, namely wpath, to combine these two approaches, using IC to weight the shortest path length between concepts. Conventional corpus-based IC is computed from the distributions of concepts over textual corpus, which is required to prepare a domain corpus containing annotated concepts and has high computational cost. As instances are already extracted from textual corpus and annotated by concepts in KGs, graph-based IC is proposed to compute IC based on the distributions of concepts over instances. Through experiments performed on well-known word similarity datasets, we show that the wpath semantic similarity method has produced statistically significant improvement over other semantic similarity methods. Moreover, in a real category classification evaluation, the wpath method has shown the best performance in terms of accuracy and F score.

II. LITERATURE SURVEY

All researches have aimed to develop and provide the generalized solution to monitor systematic way representing semantics in words using knowledge graph which can improve the efficiency of the database record and reduce the space between the data retrieval system. The major contributions to these topics are summarized below.

A. Information Retrieval by Semantic Similarity [1]

1) *Author:* Angelos Hliaoutakis, Giannis Varelas Department of Electronics and Computer Engineering, Greece
2) *Publication:* Volume 4, Issue 4, April 2016 International Journal of Advance Research.

Semantic Similarity relates to computing the similarity between conceptually similar but not necessarily lexically similar terms. Typically, semantic similarity is computed by mapping terms to an ontology and by examining their relationships in that ontology. We investigate approaches to computing the semantic similarity between natural language terms (using WordNet as the underlying reference ontology) and between medical terms (using the MeSH ontology of medical and biomedical terms). The most popular semantic similarity methods are implemented and evaluated using WordNet and MeSH. Building upon semantic similarity we propose the Semantic Similarity based Retrieval Model (SSRM), a novel information retrieval method capable for discovering similarities between documents containing conceptually similar terms. The most effective semantic similarity method is implemented into SSRM.SSRM has been applied in retrieval on OHSUMED (a standard TREC collection available on the Web). The experimental results demonstrated promising performance improvements over classic information retrieval methods utilizing plain lexical matching (e.g., Vector Space Model) and also over state-of-the-art semantic similarity retrieval methods utilizing ontologies. mputer Science and Management Studies.

B. Design and Evaluation of Semantic Similarity Measures for Concepts Stemming from the Same or Different Ontologies [2]

1) Author: Euripides G.M. Petrakis.

Semantic Similarity relates to computing the similarity between concepts (terms) which are not necessarily lexically similar. We investigate approaches to computing semantic similarity by mapping terms to an ontology and by examining their relationships in that ontology. More specifically, to investigate approaches to computing the semantic similarity between natural language terms (using WordNet as the underlying reference ontology) and between medical terms (using the MeSH ontology of medical and biomedical terms). The most popular semantic similarity methods are implemented and evaluated using WordNet and MeSH. The focus of this work is also on cross ontology methods which are capable of computing the semantic similarity between terms stemming from different ontologies (WordNet and MeSH in this work). This is a far more difficult problem (than the single ontology one referred to above) which has not been investigated adequately in the literature. X-Similarity, a novel cross-ontology similarity method is also a contribution of this work. All methods examined in this work are integrated into a semantic similarity system which is accessible on the Web.

C. Survey Of Semantic Similarity Measures In Pervasive Computing [3]

1) Author: Djamel Guessoum, Moeiz Miraoui,
2) Publication: INTERNATIONAL JOURNAL ON SMART SENSING AND INTELLIGENT SYSTEMS VOL. 8, NO. 1, MARCH 2015

Semantic similarity measures usage is prevalent in pervasive computing with the following aims: 1) to compare the components of an application; 2) to recommend and rank services by degree of relevance; 3) to identify services by matching the description of a

query with the available services; 5) to compare the current context with already known contexts. The existing works that apply semantic similarity measures to pervasive computing focus on one particular issue. Furthermore, surveys in this domain are limited to the recommendation or discovery of context-aware services. In this article, we therefore present a survey of context-aware semantic similarity measures used in various areas of pervasive computing.

D. A Survey on Semantic Similarity Measure [4]

1) *Author:* S. Anitha Elavarasi1, Dr. J. Akilandeswari Department of Computer Science and Engineering Department of Information Technology Sona College of Technology
2) *Publication:* International Journal of Research in Advent Technology, Vol.2, No.3, March 2014.

Measuring semantic similarity between concepts is an important problem in web mining and text mining which needs semantic content matching. Semantic similarity has attracted great concern for a long time in artificial intelligence, psychology and cognitive science. Many measures have been proposed. The paper contains a review of the state of art measures including path based measures information based measures, feature based measures and hybrid measures. The features, performance advantages, disadvantages and related issues of different measures are discussed. This paper makes a review of semantic similarity measures with various approaches.

E. Development and application of a metric on semantic nets [5]

1) *Author:* R. Rada Dept. of Comput. Sci., Liverpool Univ., UK. E. Bicknell
2) *Publication:* IEEE Transactions on Systems, Man, and Cybernetics (Volume: 19, Issue: 1, Jan/Feb 1989)

Motivated by the properties of spreading activation and conceptual distance, the authors propose a metric, called distance, on the power set of nodes in a semantic net. Distance is the average minimum path length over all pairwise combinations of nodes between two subsets of nodes. Distance can be successfully used to assess the conceptual distance between sets of concepts when used on a semantic net of hierarchical relations. When other kinds of relationships, like 'cause', are used, distance must be amended but then can again be effective. The judgements of distance significantly correlate with the distance judgements that people make and help to determine whether one semantic net is better or worse than another. The authors focus on the mathematical characteristics of distance that presents novel cases and interpretations. Experiments in which distance is applied to pairs of concepts and to sets of concepts in a hierarchical knowledge base show the power of hierarchical relations in representing information about the conceptual distance between concepts.

F. Proposed Approach

Implementation is the stage of the project when the theoretical design is turned out into a working system. Thus it can be considered to be the most critical stage in achieving successful new system and in giving the user, confidence that the new system will work and be effective. The implementation stage involves careful planning, investigation of the existing system and it’s constraints on implementation, designing of methods to achieve changeover and evaluation of changeover methods.

III. METHODOLOGY AND MODULES

A. Modules

1) *WPath Semantic Similarity Metric:* The knowledge-based semantic similarity metrics mentioned in the previous section are mainly developed to quantify the degree to which two concepts are semantically similar using information drawn from concept taxonomy or IC. Metrics take as input a pair of concepts, and return a numerical value indicating their semantic similarity. Many applications rely on this similarity score to rank the similarity between different pairs of concepts. Considering both advantages and disadvantages of conventional knowledge-based semantic similarity methods, we propose a weighted path length (wpath) method to combine both path length and IC in measuring the semantic similarity between concepts. The IC of two concepts’ LCS is used to weight their shortest path length so that those concept pairs having same path length can have different semantic similarity score if they have different LCS.
2) *Graph-Based Information Content:* Conventional corpus-based IC requires to prepare a domain corpus for the concept taxonomy and then to compute IC from the domain corpus in offline. The inconvenience lies in the high computational cost and difficulty of preparing a domain corpus. More specifically, in order to compute corpus-based IC, the concepts in the taxonomy need to be mapped to the words in the domain corpus. Then the appearance of concepts are counted and the IC values for concepts are generated. In this way, the additional domain corpus preparation and offline computation may prevent the

application of those semantic similarity methods relying on the IC values (e.g., res, lin, jcn, and wpath) to KGs, especially when the domain corpus is insufficient or the KG is frequently updated. Since KGs already mined structural knowledge from textual corpus, we present a convenient graph-based IC computation method for computing the IC of concepts in a KG based on the instance distributions over the concept taxonomy.

3) *Word Similarity Evaluation:* All the datasets described above contain a list of triples comprising two words and a similarity score denoting word similarity judged by human subjects. The human ratings on those word pairs have been proven to be highly replicable. This indicates that human assessment about semantic similarity between words is remarkably stable over a large time span and such datasets containing human ratings can be reliably used for evaluating semantic similarity methods. Since those datasets contain different coverage of word pairs, we use all the datasets for evaluation in order to present a more completed and objective experiment. Those datasets are used for evaluating word similarity. However, the semantic similarity metrics presented in this paper are used for concepts, rather than words. We convert those concept-to-concept semantic similarity metrics into a word-to-word similarity metrics by taking the maximal similarity score over all the concepts which are the senses of the words.

IV. CONCLUSION

Measuring semantic similarity of concepts is a crucial component in many applications which has been presented in the introduction. In this paper, we propose wpath semantic similarity method combining path length with IC. The basic idea is to use the path length between concepts to represent their difference, while to use IC to consider the commonality between concepts. The experimental results show that the wpath method has produced statistically significant improvement over other semantic similarity methods. Furthermore, graph-based IC is proposed to compute IC based on the distributions of concepts over instances. It has been shown in experimental results that the graph-based IC is effective for the res, lin and wpath methods and has similar performance as the conventional corpus-based IC. Moreover, graph-based IC has a number of benefits, since it does not requires a corpus and enables online computing based on available KGs. Based on the evaluation of a simple aspect category classification task, the proposed wpath method has also shown the best performance in terms of accuracy and F score.

BIBLIOGRAPHY

[1] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, “Free- base: a collaboratively created graph database for structuring human knowledge,” in Proceedings of the 2008 ACM SIGMOD international conference on Management of data. ACM, 2008, pp.1247–1250

[2] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann, “Dbpedia-a crystallization point for the web of data,” Web Semantics: Science, Services and Agents on the World Wide Web, vol. 7, no. 3, pp. 154 – 165, 2009, the Web of Data.

[3] J. Hoffart, F. M. Suchanek, K. Berberich, and G. Weikum, “Yago2: A spatially and temporally enhanced knowledge base from wikipedia (extended abstract),” in Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, ser. IJCAI ’13.AAAI Press, 2013, pp. 3161–3165.

[4] I. Horrocks, “Ontologies and the semantic web,” Commun. ACM, vol. 51, no. 12, pp. 58–67, Dec. 2008. [Online]. Available: http://doi.acm.org/10.1145/1409360.1409377

[5]G. A. Miller, “Wordnet: a lexical database for English,” Communications of the ACM, vol. 38, no. 11, pp. 39–41, 1995.

[5] Z. Wu and M. Palmer, “Verbs semantics and lexical selection,” in Proceedings of the 32nd annual meeting on Association for Computational Linguistics, ser. ACL ’94. Stroudsburg, PA, USA: Association for Computational Linguistics, 1994, pp. 133–138.

[6] Y. Li, Z. Bandar, and D. Mclean, “An approach for measuring semantic similarity between words using multiple information sources,” Knowledge and Data Engineering, IEEE Transactions on, vol. 15, no. 4, pp. 871–882, 2003.

[7] J. J. Jiang and D. W. Conrath, “Semantic similarity based on corpus statistics and lexical taxonomy,” Computational Linguistics, vol. cmp-lg/970, no. Rocling X, p. 15, 1997.

[8] D. Lin, “An information-theoretic definition of similarity,” in Pro- ceedings of the Fifteenth International Conference on Machine Learning, ser. ICML ’98. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1998, pp. 296–304.

[9] M. Pontiki, D. Galanis, H. Papageorgiou, S. Manandhar, and I. Androutsopoulos, “Semeval-2015 task 12: Aspect based senti- ment analysis,” in Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Association for Computational Linguistics, Denver, Colorado, 2015, pp. 486–495