International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 8 | ISSN : 2456-3307
DOI : https://doi.org/10.32628/ CSEIT183897

Survey on Semantic Similarity

Tanmay Joshi
Department of Computer Engineering, Pune Institute of Computer Technology, Pune, Maharashtra, India

ABSTRACT

Semantic similarity is the measure of similarity in the meanings represented by different terms or sentences. There are many different ways in which a statement can be expressed by using various words conveying the same meaning. Also, a single word can mean a lot of different things in various contexts. Hence, semantic similarity plays a major role in data processing, data mining and artificial intelligence applications. In order to compute semantic similarity, many different methods have been proposed by various researchers. This paper makes a review of the various measures for computation of semantic similarity.

Keywords: Semantic Web, Ontology/ Taxonomy, Natural Language Processing.

I. INTRODUCTION

Semantic similarity plays a very important role in the field of data processing, data mining and artificial intelligence. It is also useful in information management, especially in the context of environment such as semantic web where data may originate from different sources and has to be integrated in fl exible way. As a lot of data are being uploaded on the internet everyday, it is essential to know what exactly it corresponds to. There are many sources that upload some data which are very similar. Hence the measurement of similarity between has become a need of the hour for the building of efficient information retrieval systems. Broadly, there are two types of similarity, namely lexical similarity and semantic similarity:

- Lexical Similarity- It is often computed between two languages. It is the computation of overlapping vocabularies.
- Semantic Similarity- It is the similarity between words or terms which give us an idea about how close their meanings are.

Semantic similarity is basically a measure used to compute the extent of similarity between two concepts based on the likeliness of their meaning. The concepts can be sentences, words or paragraphs. The calculation of semantic similarity among various terms or words is a very basic problem in the field of lexical semantics and thus has various applications related to information retrieval like web search, document search etc.

Semantic similarity refers to the closeness of the meaning of the concerned terms. Any two terms are said to be semantically similar if their meaning points to a certain specific object or entity. It finds the distance between different concepts in semantic space in such a way that lesser the distance, greater the similarity.

The techniques used to find the semantic similarity between different words can be extended to find similarity between phrases, sentences or paragraphs. Lexical similarity is presented using different string based algorithm and semantic similarity is presented using Corpus based and Knowledge based algorithms.

Semantic similarity measures are being intensively used in various applications of knowledge based and semantic information retrieval systems for identifying an optimal match between user query terms and documents. It is also used in word sense disambiguation for identifying the correct sense of the term in the given context. Semantic similarity and semantic relatedness are two different concepts but related to each other. For example,"educational institute" and "university" are similar because their semantic contents are very similar. Also, "Google" and "Microsoft" are similar because they are both software companies. But, "mother" and "child" are related terms but are not similar since they have different meaning.

II. REVIEW OF LITERATURE

For computing semantic similarity, there are two major methods:

- Knowledge based similarity- It is a semantic similarity measure that determines the degree of similarity between words using information derived from semantic networks. Knowledge based approaches depends on handcrafted resources such as thesauri, taxonomies or encyclopedias, as the context of comparison. Most of this work depends on the linguistics of isA relations in words that could be a manually created lexicon and taxonomy.
- Corpus based similarity- It is a semantic similarity measure that determines the similarity between words according to information gained from large corpora. Corpus based approaches primarily work by extracting the contexts of the terms from giant corpora. Corpuses are often something from web pages, internet search snippets to different text repositories.

Over the years, a lot of research has been done in the field of semantic similarity, its computation and application. In 1989, Rada et al.[5] proposed path based approach in order to compute semantic similarity. A new metric called "distance" was introduced. The distance is used to compute the similarity between two nodes in a tree hierarchy suchk as WordNet. This approach gives low accuracy as the amount of information hidden between the nodes is ignored.

Certain other approaches were also introduced in which the Information Content[4] of the nodes in tree hierarchy of WordNet were evaluated. With the introduction of Information Content, better results were obtained as compared to the path based methodologies. Variations of graph-based algorithms were also applied on WordNet using a rooted weighted graph is constructed using WordNet. Most of these techniques use knowledge based data of WordNet, which has limited number of taxonomies covered in it. Also, WordNet can not be used in case of various new terms and expressions that people use every day.

Along with the knowledge based approaches, researchers have done work on various corppus based approaches also. Rohde et al.[10] introduced a vector-space method for deriving word-meanings from large corpora. Kazama et al.[11] proposed a Bayesian method for robust distributional word similarities. These methods fetch better results when semantic similarity is computed using them, but these are derived from distributional models which are symbolic and give us an idea about perception rather than action.

These methods for some cases may prove to be effective but they are also costly in terms of time as

parsing of the texts is involved. Also, the number of taxonomies covered are limited in these cases.

III. REFERENCES

[1] P. Li, H. Wang, K. Q. Zhu, Z. Wang, X. Hu, and X. Wu, "Large Probabilistic Semantic Network Based Approach to Compute Term Similarity" in *IEEE Transactions on Knowledge and Data Engineering*, Vol. 27, No. 10, October 2015

[2] Kavitha Sri.M and Hemalatha.P. "Survey on Text Classification Based on Similarity" in *International Journal of Innovative Research in Computer and Communication Engineering*, Vol. 3, Issue 3, March 2015

[3] S.S. Kulkarni and K.S.Kadam, "Development of Term Similarity Measurement Using Semantic Network Approach" in *International Journal of Advance Engineering and Research Development*, Volume 3, Issue 11, November 2016.

[4] P. Ransik. "Using Information Content to Evaluate Semantic Similarity in a Taxonomy",in *Proc. 14th Int. Joint Conf. Artif. Intell.*,1995, pp. 448-453.

[5] R. Rada, H. Mili, E. Bicknell, and M. Blettner. "Development and Application of a Metric on Semantic Nets" in *IEEE Transactions On Systems. Man. And Cybfrnetics*, Vol 19. January 1989.

[6] Q. Do, D. Roth, M. Sammons, Y. Tu and V.G.V. Vydiswaran. "Robust, Light-weight Approaches to compute Lexical Similarity" *University of Illinois, Champaign, IL, USA, Comput. Sci. Res. Tech. Rep.*, 2009

[7] D. Sánchez, M. Batet and D. Isern, "Ontology-based information content computation" *Knowledge-Based Systems* vol 24, 2011 pp. 297-303.

[8] E. Banerjee and T. Pedersen. "An adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet", in *Proc. 3rd Int. Conf. Comput. Linguistics Intell. Text Process*, 2002, pp. 136 – 145.

[9] D. Bollegala, Y. Matsuo, and M. Ishizuka. "A Web Search Engine-Based Approach to Measure Semantic Similarity between Words", in *IEEE Transactions On Knowledge And Data Engineering*, Vol. 23, No. 7, July 2011.

[10] D. L. T. Rohde, L. M. Gonnerman, and D. C. Plaut, "An improved model of semantic similarity based on lexical co-occurrence," in *Commun. ACM*, vol. 8, pp. 627 - 633, 2005.

[11] J. Kazama, S. D. Saeger, K. Kuroda, M. Murata, and K. Torisawa, "A Bayesian method for robust estimation of distributional similarities" in *Proc. 48th Annu. Meeting Assoc. Comput. Linguistics*, 2010, pp. 247 - 256.

Cite this article as :

Tanmay Joshi, "Survey on Semantic Similarity ", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 3 Issue 8, pp. 403-405, November-December 2018.
Journal URL : http://ijsrcseit.com/CSEIT183897