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/CSEIT183896

Context-Based Semantic Similarity and Document Retrieval

Tanmay Joshi*[1], Prof. A. G. Phakatkar

*[1,2]Department of Computer Engineering, Pune Institute of Computer Technology, Pune, Maharashtra, India

ABSTRACT

Text based methods are extensively used in information retrieval on web. 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. It is a challenging task to make the system understand what exactly the word or statement means and in which context. Hence, it is an important task to find out the context of the words so as to effectively understand what the user intends to mean which in turn can be used in a variety of NLP and text processing applications. A method is proposed in order to find out the similarity between two words and use the results in order to retrieve documents.

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

I. INTRODUCTION

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. Different words can mean the same thing as well as same words can mean different things in any given language. This is called context. For example, the word 'apple' may mean a software organization in one context whereas it may mean a fruit or a tree in some other context. It depends on human perception. But computers do not have perception. Hence computation of context of terms is an issue in the field of lexical semantics. Also it is essential to compute the closeness of meanings of various words. Hence the measurement of similarity between has become a need of the hour for the building of efficient information retrieval systems.

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. Similarity functions are used to compute the value of semantic similarity, which is between 0 and 1; higher the value, higher is the similarity between the concerned terms.

For example, "educational institute" and "university" are similar because their semantic contents are very similar. Another example, "Google" and "Microsoft" are similar because they are both software companies. However, "car" and "journey" are not semantically similar but they both are related with each other because "car" is a transport means for the activity "journey". On an isA taxonomy semantic similarity can be defined as measuring distance between two terms.

There are two broad approaches to compute semantic similarity namly Knowledge based approaches and Corpus based approaches. The 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. The 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.

Both the above-mentioned approaches have several disadvantages. The knowledge based approaches highly rely on knowledge bases like Wordnet which have limited coverage of taxonomies in isA relationship while in case of corpus based approaches, there is a high possibility of getting biased results because of indexing and ranking used in search engines. Also, one of the main disadvantage of these approaches is that they do not take into consideration, the context of words.

In order to cover these disadvantages, Microsoft has developed a large scale probabilistic semantic network called "Probase". Probase is basically a large dataset extracted from search logs of Bing search engine. It has advantages over the knowledge based and corpus based methods because, a large number of taxonomies are covered and the whole data is organized in isA relationship which makes it easier to find out various contexts of words. We have made use of Probase data in our system.

The organization of this document is as follows. In Section 2 (**Survey**), a survey related to semantic similarity is done to give an idea of various approaches. In Section 3 (**System Overview**), the implemented system is described. In Section 4 (**Methodology**), a detail explaination of the working of our system is given. In Section 5 (**Results**), the results obtained by our system are given. In Section 6 (**Conclusion**), concluding remarks of the paper are made.

II. SURVEY

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 such 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. SYSTEM OVERVIEW

The function of the system is to compute semantic similarity of two terms based on their contexts and retrieve the relevant documents depending upon the contexts.

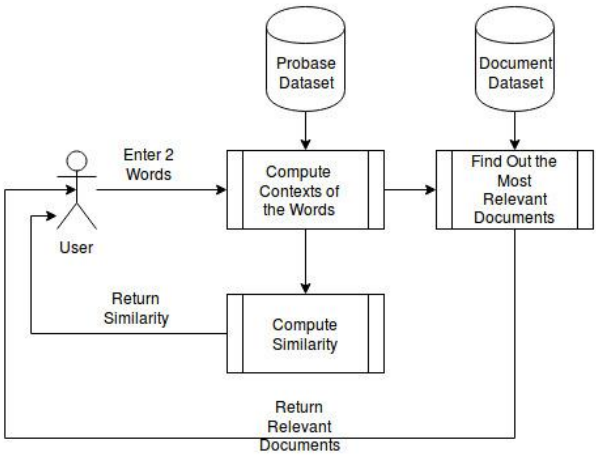

Fig. 1: System Architecture

In the system, the user has to enter two terms. Based on the contexts of the terms, similarity between them is computed. Then document classification is done based on the similarity of the words and their contexts.

A. Dataset Used

For the computation of similarity, data from Probase dataset is used, because it provides more number of taxonomies as compared to other approaches like knowledge based or corpus based.

The Probase data is organized in the form of isA relationship between words along with the count of their occurrence. Information is in the form of (c,e,W), where c is concept (hypernym), e is entity (hyponym), W is the count of occurrence of (c,e). For example, (state, new york, 8125), (fruit, apple, 6315) etc.

IV. METHODOLOGY

The first task of our system is to compute the contexts of the entered words. For the computation of contexts of the words, we make use of Probase dataset which is described in earlier section of this paper.

From Probase data, we can find out the contexts of the words by figuring out the co-occuring words. For example if the user has entered input words as {microsoft, apple}, then the contexts of the word 'microsoft' is {(company, 6189), (vendor, 898), (client, 489), (firm, 461)…} and that of apple is {(fruit, 6315), (company, 4353), (food, 1152), (brand, 764)…}. The number corrosponding to the context is the number of times the two words have co-occured i.e. microsoft and company have co-occured 6189 times.

In this way, we compute the contexts of individual words and then, the common contexts are found out. For example, company is the common context in the above example. Similarly all the common contexts of the two words are figured out and then vectors of the common contexts for each word are generated. For example, if {company, organization, software provider} are the common contexts of 'apple' and 'microsoft', then the vector for 'apple' would be {(company, 4353), (organization, 3512), (software provider, 3175)} and that of 'microsoft' would be {(company, 9189), (organization, 4627), (software provider, 2716)}.

In order to compute the semantic similarity, cosine similarity function is applied on the vectors and the desired similarity is returned by the system.

The second output of the system is the retrieval of most relevant documents from a set of the documents. For this purpose, the term frequency and inverse document frequency corrosponding to the input

words and top 5 contexts is computed and the documents are sorted according to their relevency.

V. RESULTS

The following are some results of semantic similarity between two words given by our system.

- Apple and Pear- 99.16% similarity and some of the contexts are fruit, corp, food, delicious fruit etc.
- Delhi and Mumbai- 88.19% similarity and some contexts are city, big indian city, metropolitian area, famous city, indian city etc.
- Google and Apple- 96.58% similarity and some of the contexts are company, vendor, top brand, technology firm, firm etc.

VI. CONCLUSION

As our system uses Probase dataset which covers a wide variety of taxonomies, it gives more efficient results as compared to the systems which use knowledge based or corpus based approaches like WordNet. Also, our system is capable of computing similarity between noun words like company names, city names etc.

The use of contexts give a more human interpretable way of understanding what exact sense do the words have and the contexts can be applied to build various applications like the retrieval of relevant documents.

VII. REFERENCES

The heading of the References section must not be numbered. All reference items must be in 10 pt font. Please use Regular and Italic styles to distinguish different fields as shown in the References section. Number the reference items consecutively in square brackets (e.g. [1]).

[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, Prof. A. G. Phakatkar, "Context-Based Semantic Similarity and Document Retrieval ", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 3 Issue 8, pp. 398-402, November-December 2018.
Available at
doi : https://doi.org/10.32628/CSEIT183896
Journal URL : http://ijsrcseit.com/CSEIT183896