Computing Semantic Similarity of Concepts in Knowledge Graphs using Graph-based Information Content

[1]Miss. Vijayalaxmi C.Deshmukh., [2] Dr. Dinkar. S. Bhosale
[1]Student, M E (Computer Science & Engineering.),[2]Professor at Dept Computer Science & Engineering.
[1]Ashokrao Mane Group Of Institutions ,vathar, Shivaji University, Kolhapur.
[2]Ashokrao Mane Group Of Institutions ,vathar,Shivaji University, Kolhapur.

Abstract: This paper displays a method for measuring the semantic similarity between concepts in Knowledge Graphs (KGs) such as WordNet and DBpedia. Early work on semantic similarity methods has 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 methods, 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 the 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 public word similarity datasets, we note that the wpath semantic similarity method has produced a statistically meaningful improvement over other semantic similarity techniques. Moreover, in a real category classification evaluation, the wpath method has shown the best review concerning accuracy and F score.

***Index Terms*– semantic similarity, knowledge base, classification, wpath, Information Content, graph, wordnet**

I. INTRODUCTION

Measuring semantic similarity of concepts is a crucial component in many applications. Semantic similarity method is combining path length with Information Content (IC). The basic approach is to use the path length between concepts to express their difference, while to use IC to consider the commonality between concepts. In this work presents a method for measuring the semantic similarity between concepts in Knowledge Graphs (KGs) such as WordNet and DBpedia. Traditional work on semantic similarity methods has focused on either the structure of the semantic network between concepts (e.g., path length and depth) or just on the (IC) of concepts. To deal with this introduce a semantic similarity method to combine these two methods, using IC to weight the shortest path length among concepts. General corpus-based IC is computed from the relationships of concepts over the textual corpus, which is required to prepare a domain corpus containing explained concepts and has high computational cost. As instances are already extracted from the textual corpus and annotated by concepts in KGs, graph-based IC is proposed to compute IC based on the distributions of concepts over instances. The wpath method results in a statistically significant improvement of correlation between computed similarity scores and human judgments.

II. PROBLEM STATEMENT:

The main idea of the recommended method is to encode both the structure of the theory of taxonomy and the statistical information of concepts. Furthermore, to adapt corpus-based IC approaches to structured KGs, graph-based IC is proposed to calculate IC based on the number of concepts over instances in KGs. Consequently, using the graph-based IC in the semantic similarity method can represent the specificity and hierarchical structure of the concepts in a KG. The system uses the semantic similarity method for measuring semantic similarity between concepts in KGs and computes graph-based IC of concepts based on KGs.

III. SYSTEM ARCHITECTURE

WPath Semantic Similarity Metric The knowledge-based semantic similarity metrics discussed and mainly developed to quantify the degree to which two concepts are semantically related using information drawn from concept taxonomy or IC. Metrics take as input a pair of ideas and return a numerical value showing their semantic similarity. Many applications rely on this similarity score to rank the similarity between different pairs of concepts. Take a fragment of concept taxonomy, given the concept pairs of the applications require similarity metrics to give higher similarity value to because the concept beef and concept lamb are kinds of meat while the concept octopus is a kind of seafood.

One of the disadvantages of conventional knowledge-based approaches (e.g., path or lch) in addressing such task is that the semantic similarity of any two concepts with the same path length is the same (uniform distance problem). As illustrated based on the path and lch semantic similarity methods is the same as because those concept pairs have equal shortest path length. Some knowledge-based approaches tried to solve the drawback by including depth information in concept taxonomy.

Considering that the upper-level concepts are more general than the lower level concepts in the hierarchy, those approaches use the depth of concepts to give higher similarity value to those concept pairs which are located deeper in taxonomy. For example, the similarity is higher than the similarity of based on semantic similarity method of wup and li, because of the concept lamb and the concept beef are located deeper in the concept of taxonomy (lamb and beef are sub-concepts of meat). In order to determine the equal path length and depth problem, some knowledge-based approaches (e.g., res, lin, or jcn) proposed to include IC because different concepts usually have different IC values (e.g., the IC of meat is 6.725, and the IC of food is 6.109) so that the is different from sim octopus; shellfish. General concepts have lower informative ness thus have a lower value of IC, while more specific concepts would have a higher value of IC. For example, the IC of meat is higher than the IC of food because meat is a sub-concept of food. The idea of using IC to compute semantic similarity is that the more information two ideas share in general, the more similar they are. Using the IC of the LCS alone in the res method can realize the common data that two concepts share, however, the problem is that the similarity of any two concepts with the same LCS is the same. For example, based on res semantic similarity, although the concept pairs beef; lamb and octopus shellfish have different similarity scores, the similarity scores of concept pairs meat; seafood and beef; octopus, beef; coffee and food; coffee are the same because the LCS of the concept pairs are concept food and matter. Other methods (e.g., lin or jcn) tried to solve the drawback by including the IC of concepts being compared.

The wpath semantic similarity method is illustrated as follows.

$$sim * wpath(Ci, Cj) = \frac{1}{1 + length(Ci, Cj) * K^{ic(Clcs)}}$$

Where k € (0; 1) and k = 1 means that IC has no contribution in shortest path length. The parameter k represents the contribution of the LCS's IC which indicates the common information shared by two concepts.

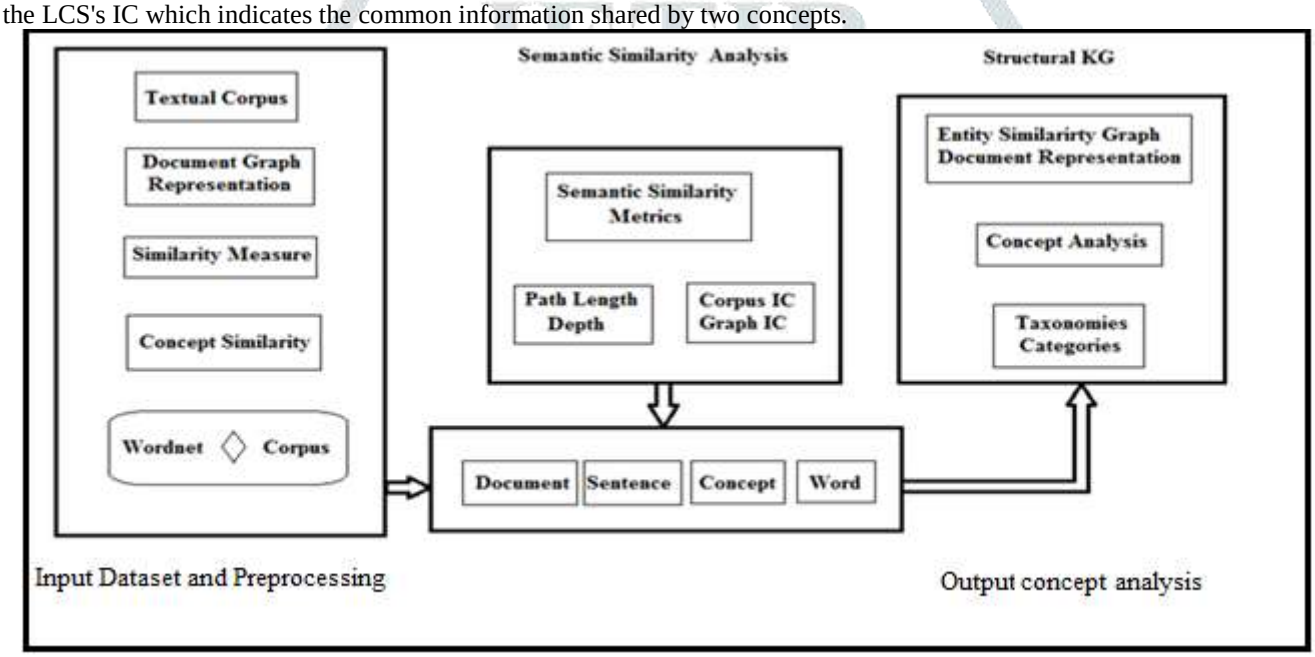

Figure.1 Architecture of Computing Semantics Similarity

In this figure 1 input dataset used as text corpus and web page metadata, Information Content (IC) is computed based on frequency counts of concepts appearing in a textual corpus. Each occurrence of a more specific concept also implies the occurrence of the more general ancestor concepts. Then document graph representation for KG is used and KG is directed designated graph, G = (V, E, £) where, V is a set of nodes, E is a set of edges connecting those nodes, t is a function V x V →E that defines all triples in G. to be accessed and applied as concept taxonomy in KGs by converting the conventional representation of Word-Net into novel linked data representation. Present an automatic approach to the development of BabelNet, a very large, wide coverage multilingual semantic network. Key to our approach is the combination of lexicographic and encyclopedic knowledge from WordNet and Wikipedia. Similarity Measure is done with knowledge-based access, it measure the semantic similarity between concepts using semantic information contained in KG. Here similarity is measured by a similarity function. Measuring semantic similarity between ideas is an essential problem in web mining and text mining which needs semantic content matching. The semantic similarity has brought great concern for a long time in artificial intelligence, psychology and cognitive science. Also concept similarity is calculated and measured by TF (term frequency) and IDF (inverse Document frequency). Concept similarity is usually based on taxonomical relations between concepts such as WordNet taxonomy and DBpedia ontology class. In this, the lexical database WordNet has been imagined as a general semantic network of the lexicon of English words. WordNet can be viewed as a concept taxonomy where nodes denote WordNet synsets describing a set of words that share one common sense (synonyms), and edges denote hierarchical relations of hypernym and hyponymy (the relation between a sub-concept and a super-concept) between synsets.

Semantic similarity metrics used for weighing or ranking similar concepts based on concept taxonomy. In such way, semantic similarity methods could be applied in KGs for concept-based entity retrieval or question answering, where those entities that contain types having a similar meaning to query concepts would be retrieved. Semantic Similarity relates to computing the similarity between conceptually similar but not necessarily lexically similar terms. Typically, semantic similarity is computed by mapping terms and by examining their relationships. We investigate approaches to computing the semantic similarity between natural language terms (using WordNet). Semantic similarity metrics is the mapping of the semantic distance between concepts utilizing hierarchical associations. Semantic similarity between two concepts is then proportionate to the length of the path connecting the two concepts. Finally combination of knowledge-based methods with the corpus-based methods is applied and result calculates for similarity metrics.

IV. RESULT ANALYSIS

a. web based dataset input analysis

Concerning graph-based IC measures, it is observed that intrinsic computation approaches which calculate the information content based on the number of concept hyponyms are clearly more accurate than corpora approaches (0.87 vs. 0.73). This refers to the fact that corpora dependency seriously frustrate the applicability of classic IC measures. Feature-based methods present a closer resemblance to those presented by structure-based measures (0.81-0.84).

This refers to the fact that they rely on concept features (synsets, features or non-taxonomic relationships) which have secondary importance in ontologies and for that reason the approaches are based on partially modeled knowledge. As a consequence, those measures need more research to outperform the approaches based on edge-counting measures. For hybrid-based measures, it is observed that the approach offers the highest accuracy (0.87) even though it is a complex approach which exploits a relative depth and relying on weighting parameters.

Method	Precision	Recall	F-measure	Accuracy
path	0.658	0.736	0.680	0.793
ICH	0.656	0.704	0.662	0.78
Wpath	0.664	0.741	0.689	0.800
*Wpath	0.670	0.750	0.700	0.872

Table 1: Accuracy analysis

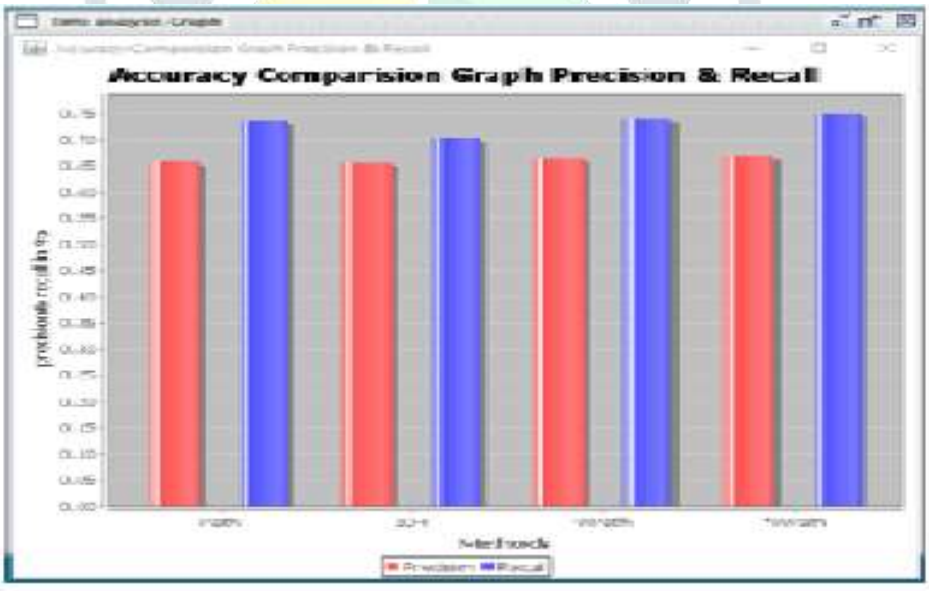

Figure 2: Accuracy analysis precision and recall

Figure 2 displays the accuracy analysis of the similarity between different inputs using precision and recall. Performance of the system is evaluated based on the precision and recall values. It shows how a given query obtains much precision and recall

Precision = TP= (TP + FP)
Recall = TP= (TP + FN)
Accuracy = TP= (TP + FP + FN)

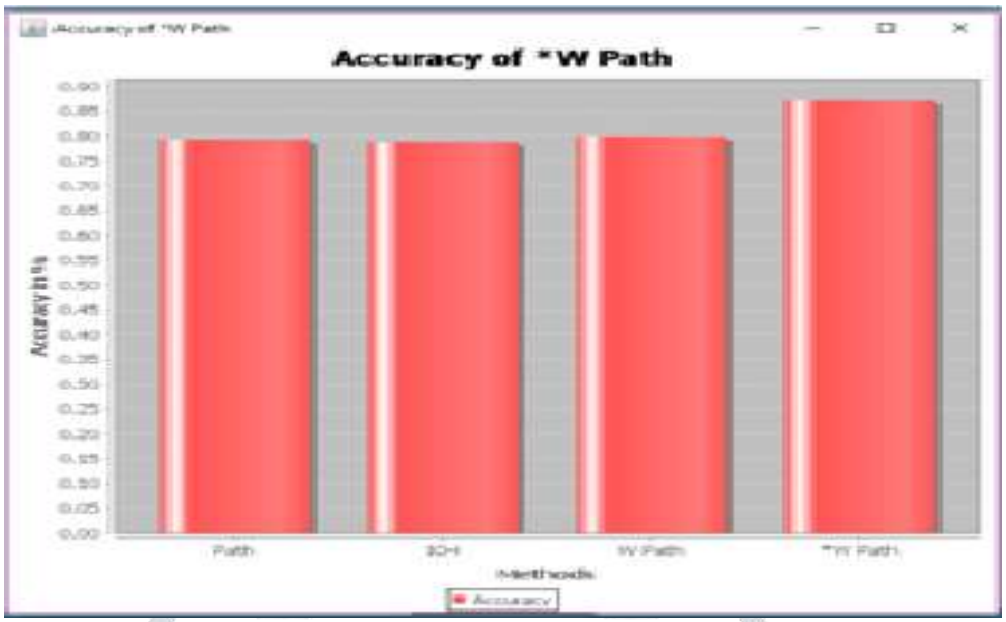

Figure 3: Accuracy analysis

Figure 3 displays the accuracy analysis of the similarity between different inputs. The accuracy of the output depends upon precision and recall.

b. Text based dataset input analysis

In figure 4 text data used as input and further process for similarity between various input files. Document similarity (or distance between texts) is one of the central themes in Information Retrieval. How humans usually define how similar document? Usually, documents treated as similar if they are semantically close and express similar concepts. On other hand \similarity" can be used in the context of parallel detection.sim2(x; y; method) - calculates the similarity between each row of matrix x and each row of matrix y using the given method. psim2(x; y; method) - calculates the parallel similarity between rows of matrix x and corresponding rows of matrix y using the given method.dist2(x; y; method) - calculates distance/dissimilarity between each row of matrix x and each row of matrix y using the given method. pdist2(x; y; method) calculates parallel distance/dissimilarity between rows of matrix x and corresponding rows of matrix y using the given method.

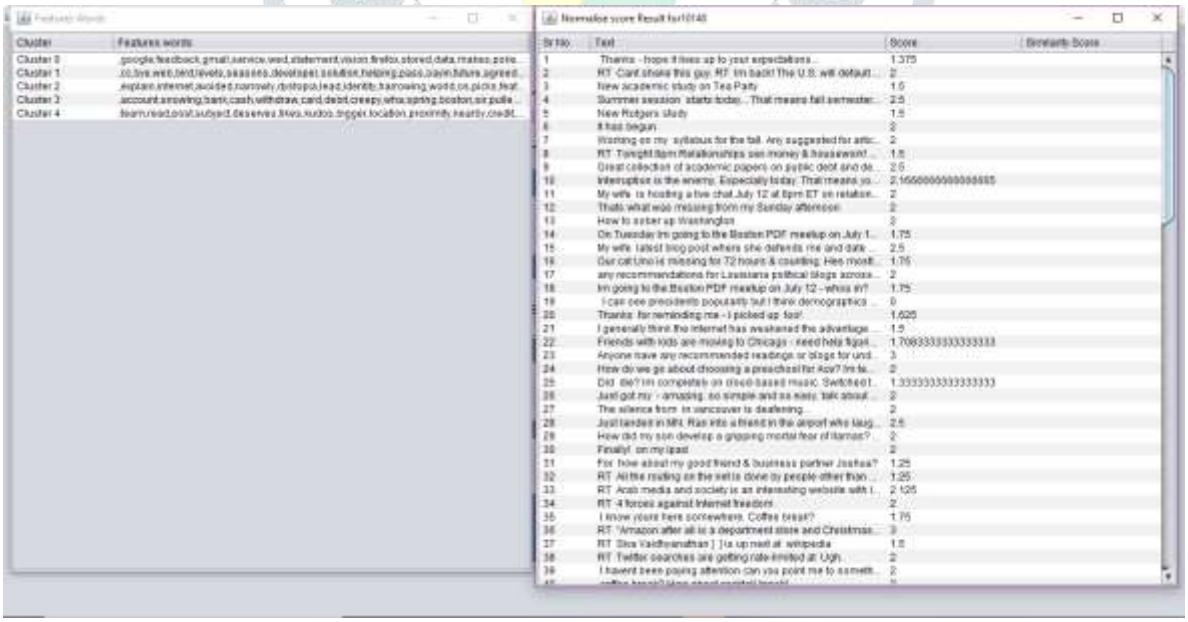

Figure 4: Similarity score text-based dataset

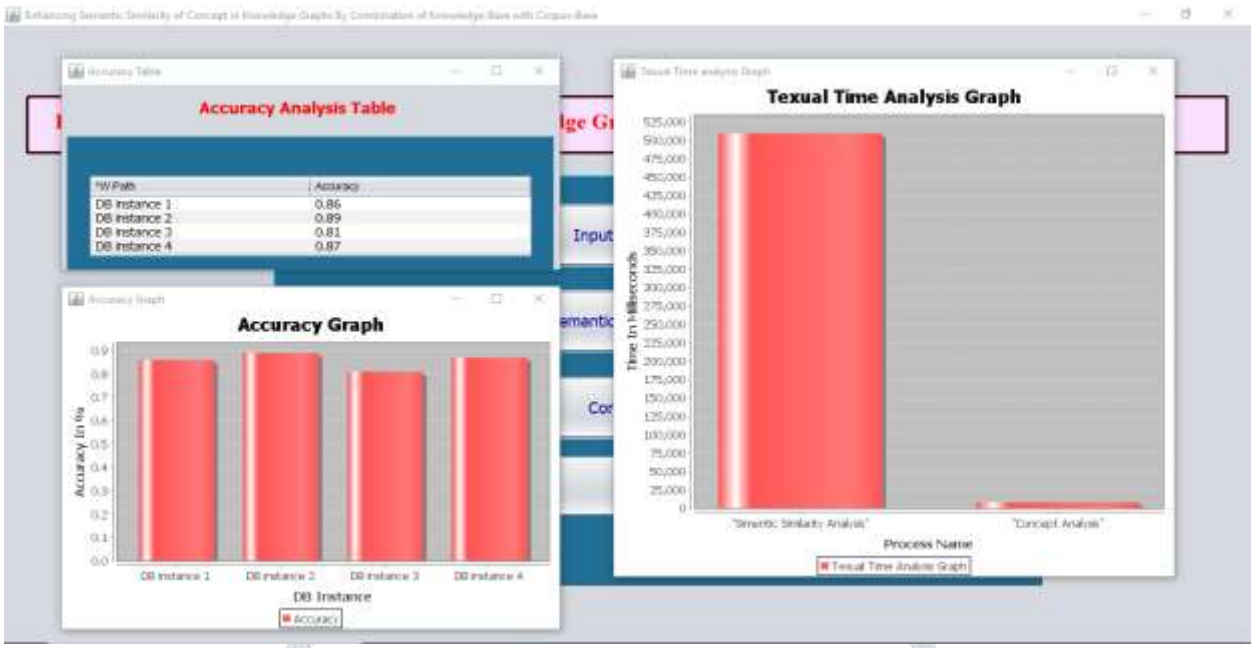

Figure 5: Output analysis text-based dataset

Figure 5 used to display accuracy and time analysis over the text-based dataset.

V. CONCLUSION

Measuring semantic similarity of ideas is a crucial component in many papers which has been presented in the introduction. In this paper the proposed semantic similarity method connecting path length with IC. The primary idea is to use the path length within concepts to draw their variety, while to use IC to consider the commonality between concepts. The test results prove that the proposed method has provided the statistically significant increase over web corpus (html, xml and txt) using semantic similarity methods. The structure of semantic web will open up the knowledge and workings of human kind to meaningful analysis by software tool, providing a new class of *wpath by which can live, work and learn together.

Furthermore, graph-based IC is proposed to compute IC based on the distributions of concepts over instances. It has been confirmed in preliminary results that the graph-based IC is useful and has similar performance to the conventional corpus-based IC. Moreover, graph-based IC has some advantages, since it does not needs a corpus and enables online computing based on prepared KGs. Based on the evaluation of a simple aspect class classification task, the proposed method has also shown the best production regarding accuracy and F-score.

REFERENCES

1. A. Miller, \WordNet: A lexical database for english," Commun. ACM, vol. 38, no. 11, pp. 39{41, 1995.
2. R. Navigli and S. P. Ponzetto, \Babelnet: The automatic construction, eval-uation, and application of a wide-coverage multilingual semantic network,"Artif. Intell., vol. 193, pp. 217{250, 2012.
3. E. Hovy, R. Navigli, and S. P. Ponzetto, \Collaboratively built semi-structured content and artificial intelligence: The story so far," Artif. Intell., vol. 194, pp. 2{27, 2013.
4. R. Navigli, \Word sense disambiguation: A survey," ACM Comput. Surveys, vol. 41, no. 2, 2009, Art. no. 10.
5. Angelos Hliaoutakis, Giannis Varelas \Information Retrieval by Semantic Similarity Department of Electronics and Computer Engineering,\ International Journal of Advance Research Greece Publication: Volume 4, Issue 4, April 2016.
6. Euripides G.M. Petrakis \Design and Evaluation of Semantic Similarity Measures for Concepts Stemming from the Same or Di_erent Ontologies."
7. Djamel Guessoum, Moeiz Miraoui Survey Of Semantic Similarity Measures In Pervasive Computing INTERNATIONAL JOURNAL ON SMART SENSING AND INTELLIGENT SYSTEMS VOL. 8, NO. 1, MARCH 2015
8. S. Anitha Elavarasi1, Dr. J. Akilandeswari A Survey on Semantic Similarity Measure Department of Computer Science and Engineering Department of Information Technology Sona College of Technology Publication: International Journal of Research in Advent Technology, Vol.2, No.3, March 2014
9. R. Rada Dept. of Comput. Sci., Liverpool Univ., UK. E. Bicknell Development and application of a metric on semantic nets IEEE Transactions on Systems, Man, and Cybernetics (Volume: 19, Issue: 1, Jan/Feb 1989)
10. R. Mihalcea, C. Corley, and C. Strapparava, \Corpus-based and knowledge-based measures of text semantic similarity," in Proc. 21st Nat. Conf. Artif. Intell., vol. 6, 2006, pp. 775{780.
11. K. W. Church and P. Hanks, \Word association norms, mutual information, and lexicography," Comput. Linguist., vol.

16, no. 1, pp. 22{29, Mar. 1990.
12. R. Gligorov, W. ten Kate, Z. Aleksovski, and F. van Harmelen, \Using google distance to weight close ontology matches," in Proc. 16th Int. Conf.World Wide Web, 2007, pp. 767{ 776.
13. T. K. Landauer and S. T. Dumais, \A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge," Psychological Rev., vol. 104, no. 2, 1997, Art. no. 211.
14. E. Gabrilovich and S. Markovitch, \Computing semantic relatedness using wikipedia-based explicit semantic analysis," in Proc. 20th Int. Joint Conf.Artif. Intell., 2007, vol. 7, pp. 1606{1611.
15. T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, \Distributed representations of words and phrases and their compositionality," in Ad-vances in Neural Information Processing Systems 26, C. Burges, L. Bottou,M. Welling, Z. Ghahramani, and K. Weinberger, Eds. Red Hook, NY, USA:Curran, 2013, pp. 3111{ 3119.
16. J. Pennington, R. Socher, and C. D. Manning, \Glove: Global vectors for word representation," in Proc. Empiricial Methods Natural Language Pro-cess., 2014, vol. 12, pp. 1532{1543.
17. Ganggao Zhu and Carlos A. Iglesias \Computing Semantic Similarity of Con-cepts in Knowledge Graphs" IEEE Tranc. On Knowlwdge and data Engg.,vol.29, no. 1, Jan 2017.
18. F. Hill, R. Reichart, and A. Korhonen, \Simlex-999: Evaluating semantic models with (genuine) similarity estimation," arXiv:1408.3456, 2014.
19. D. Sanchez, M. Batet, D. Isern, and A. Valls, \Ontology-based semantic similarity: A new feature-based approach," Expert Syst. Appl., vol. 39, no.9, pp. 7718{7728, 2012.
20. A. Tversky, \Feaures of similarity," Psychological Rev., vol. 84, pp. 327{352,1977.
21. T. Pedersen, S. Patwardhan, and J. Michelizzi, \WordNet::similarity:Measuring the relatedness of concepts," in Proc. Demonstration Papers at HLT-NAACL, 2004, pp. 38{41. [Online]. Available:http://dl.acm.org/citation.cfm?id=1614025.1614037
22. J. H. Steiger, \Tests for comparing elements of a correlation matrix," Psy-chological Bulletin, vol. 87, no. 2, 1980, Art. no. 245.