Improved Corpus-based Word and String Similarity for Semantic Text Similarity

K.Ananthi, Assistant Professor
Department of Computer Science and Business Systems
Sri Eshwar College of Engineering,
Kinathukadavu- 641 202

Dr.M.Balakrishnan, Associate Professor
Department of Artificial Intelligence and Data Science
Dr.Mahalingam College of Engineering and Technology,
Pollachi-642003

Abstract

Using a corpus-based measure of semantic word similarity and a normalized and modified variation of the Longest Common Subsequence (LCS) string matching technique, we describe a method for assessing the semantic similarity of texts. Large documents (such as text categorization and information retrieval) or specific words were the primary focus of earlier research on this topic (e.g. synonymy tests). Our work focuses on assessing the semantic similarity of short texts because a large amount of the current content, both online and off, consists of shorter text samples (e.g., summaries of scientific documents, picture captions, and product descriptions). The main objective of our analysis is to determine how much two lines or short paragraphs resemble one another. The proposed approach may be helpful.

Keywords—wordnet, semantic similarity, similarity metric, graph theory, corpus based similarity

INTRODUCTION

Among the primary study areas in text mining are concept extraction, natural language processing, information extraction, search and information retrieval, clustering, document categorization, web mining, and information extraction. Applications from many areas frequently have to search for related documents given a query document. In order to use this feature, you must have two things: I a concept of document similarity, and (ii) an effective way to locate pertinent documents throughout potentially vast corpora. The word-level approaches of Masoud Reyhani Hamedani et al. [12] are susceptible to problems brought on by polysemy and synonymy words with similar meanings since they infer semantics from text without using explicit knowledge. The elements in the knowledge base network and their connections are precisely specified as a semantic graph. Typically, two texts are compared using a straightforward lexical matching method, and the degree of similarity between the texts is indicated by a similarity score that is calculated based on the number of lexical units that appear in both input segments. Stemming, stop-word removal, part-of-speech tagging, longest subsequence matching, as well as different weighting and normalization variables, have all been added into the proposed method to make it more effective. One of the most common methods for figuring out how semantically related two documents are is the Latent Semantic Analysis (LSA) approach. Another piece of work that is comparable is the semantic similarity approach, which made use of Word Net-based methods. Quantifying the links between entities and obtaining semantically pertinent information are both made possible through the use of knowledge-based entity models and graph-based approaches.

Techniques for managing text, ranging from single words to full databases of Soumyajit Ganguly and Vikkram Pudi et al [11] documents, are needed in order to convert text into a structured, numerical representation and apply analytical algorithms. Users have access to a wide range of tools for conducting relevant information searches, including keyword searching, topic- and subject browsing, and other methods that can help users locate pertinent material fast. Users can access a collection of pertinent documents using index search algorithms. Weiguo Zheng et al [13] have found that these search strategies aren't always enough. A significant difficulty is acquiring new knowledge and retrieving the meaning of text documents and connecting it to existing knowledge. The current work focuses on discovering relevant

information or knowledge components in text documents, text databases, log files, and contracts with the text similarity.

The goal of information or text extraction is to find features that have been assessed. Pre-processing is required to extract the contents. By utilizing unique text document features, SaschaRothe et al. [14] retrieved text. The content to be extracted from the papers is frequently the target. The author S. S. Sonawane et al. [19] defines the graph structures for semantic queries with nodes, edges, and attributes is depicted. Graph provides a way to quantify relationships between entities and offers a knowledge-based entity model to extract the semantically associated information. The author Lingling Meng et al. [16] described the similarity score as comparing the most similar terms in text documents made up of relevant score Every term has a unique dimension assigned to it, and each document is represented by a vector, with the value of each dimension corresponding to the frequency with which each phrase appears in the documents. The degree of similarity between the text documents is determined by text similarity-based approaches, according to B. Gipp et al. [20]. The similarity measure computes the degree of similarity between the documents. For analysis and effective information extraction, similarity between text documents is essential.

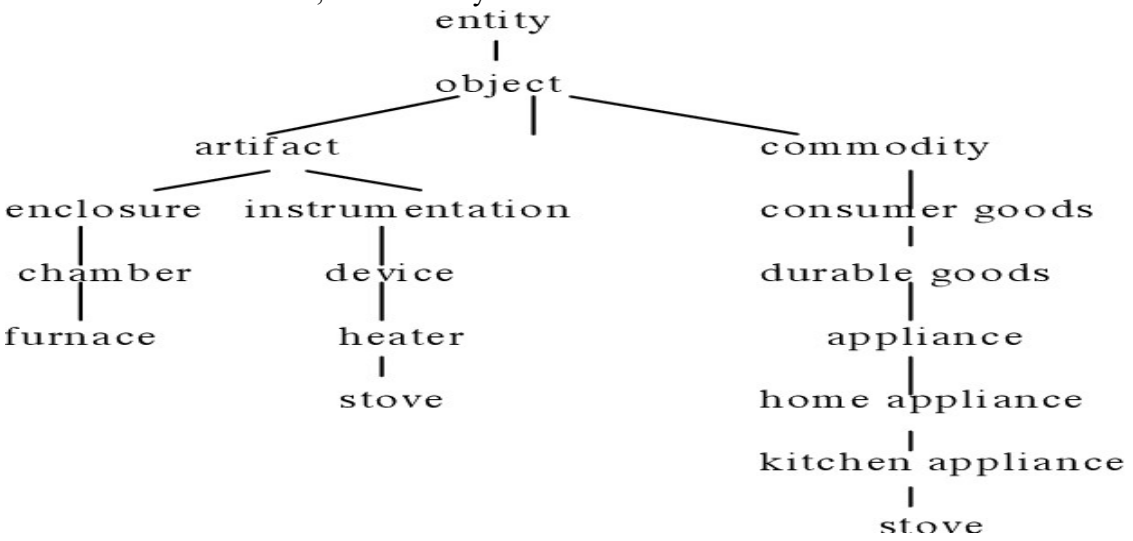

Fig.1. Semantic Similarity

Nitesh Pradhan et al's [15] description of graph based document models explains how they can produce knowledge about the relationships between the items.

LITERATURE SURVEY

Traditionally, semantic similarity metrics have been calculated between words or concepts as opposed to comparing text segments made up of two or more words. The emphasis on word-to-word similarity measurements is probably due to the accessibility of tools that particularly encode associations between words or concepts (like WordNet) and the range of testbeds that permit their evaluation (like TOEFL or SAT analogy/synonym tests). The majority of this field's research has been on applications of the conventional vectorial model, which is hardly ever extended to n-gram language models. This is because converting a word-based semantic similarity metric into a text-to-text measure of similarity may be challenging. Our goal is to automatically determine a score that measures the similarity of two inputs in terms of semantics.The eight various knowledge- and corpus-based measures of word semantic similarity are described in detail in the next section. In addition to word similarity, word

specificity is included. Word similarity measurements between abstract concepts can therefore be less significant, whereas semantic matches between specific phrases (such as collie and sheepdog) can be given more weight (e.g. get and become). Despite the fact that the depth of the semantic hierarchy already serves as a measure of the specificity of words in some ways, we are reinforcing this trait with a corpus-based measure of word specificity that is based on distributional data gathered from sizable corpora. Using a Word2Vec-based method, Yoo Kyung et al. [4] present a new topology that more exactly captures the study fields. Christian Paul et al. knowledge graph .'s entities [5] investigate how Semantic Document Expansion improves the annotations with relational information across two categories of exploited knowledge depending on the type of edge that is travelled. The text similarity algorithm is used to calculate how "close" two pieces of text are in terms of surface similarity or similarity in meaning. The topics actually refer to how apparently similar two writings are. The term "resemblance" also refers to lexical similarity.

When structured data is published using linked data, as described by Rouzbeh Meymandpour et al. [6], the semantic similarity between two concepts, entities, phrases, sentences, or documents is reflected in the relationship between their respective meanings. The measuring of similarity can be done using set-based indices like the Jaccard and Dice co-efficient.

The similarity metrics given by Majid Mohebbi et al [7] are constrained by the fact that they only take into consideration the most similar or all similar terms in the other phrase, according to a number of prior unsupervised investigations. Graph theory's Maximum Matching concept is used to create a new similarity metric that more effectively compares texts. An unsupervised knowledge-based approach that considers precise word-to-word similarities rather than just the overall or most significant similarities across sentences when assessing the semantic similarity of texts.

Zia Ul-Qayyum et al method .'s [8] extracts the sentence pairings in order to measure the semantic similarity. Both a word specificity measure and word-to-word similarity measurements are employed. Both text-based similarity and link-based similarity are used to calculate similarity. If there are more common terms between two papers, which are represented as an n-dimensional vector, the documents are assessed as being more similar. The dimension's value calculates similarity and gives the term's weight. The content is ignored as the similarity is calculated via a citation graph analysis. The hybrid technique is used to compute the semantic similarity between the papers.

The computation of semantic similarity between texts using knowledge graphs and improving similarity accuracy utilising graphs are some of the constraints of the current approach.

PROPOSED WORK

This section explains how to calculate semantic similarity and offers a recommendation for a method that is based on a graph.

a.Similar Texts in Text

The eight various corpus-based and knowledge-based metrics of word semantic similarity are explained in depth in the section that follows. We take into account both the specificity and similarity of phrases, allowing us to emphasise a semantic match between two unique words (like collie and sheepdog) while downplaying the similarities between generic ideas (e.g. get and become). We are enhancing this factor by developing a corpus-based measure of word specificity that is based on distributional data gathered from sizable corpora, even though the specificity of words is already in part determined by their position in the semantic hierarchy.

B. Corpus-based Measures

The collection opens with Baker's assessment of the need for creating a successful corpus-based methodology for finding the unique characteristics of the language of translation. According to her, the goal of this endeavour is not just to clarify the nature of the "third code" in general but, more crucially, to understand the unique restrictions, requirements, and motivations that affect the translation of and underlying its particular language. In his second piece, Shlesinger examines the drawbacks and potential advantages of corpus-based interpreting research. Instead than just being a specific sort of translation, interpreting is seen as an exceptional "mode of interlingual processing [...] shaped by its own goals, pressures, and environment of production." Two approaches to using corpora that Shlesinger examines.

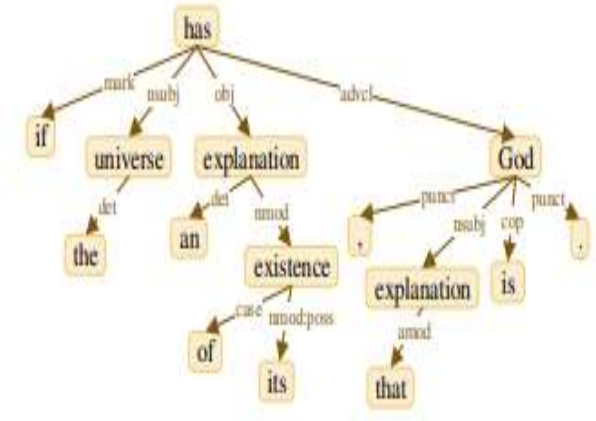

Fig.2. Graph structure

C. Latent Semantic Analysis

The following three studies—Foltz, Kintsch, and Landauer (1998/this issue), Rehder et al. (1998/this issue), and Wolfe et al. (1998/this issue)—exploit a novel theory of knowledge induction and representation (Landauer and Dumais, 1996, 1997) that provides a method for determining the similarity of meaning of words and passages by analyzing large text corpora. After processing by Latent Semantic Analysis (LSA), the words used in a significant sample of machine-readable language, as well as any group of these words—such as a sentence, paragraph, or essay—whether taken from the original corpus or new, are represented as points in a very high (e.g. 50-1,500) dimensional "semantic space." The mathematical method of singular value decomposition, which is similar to factor analysis and is the basis of LSA, is used to decompose matrices.

By accounting for word specificity, we give more weight to the predicted similarity between two particular keywords and less weight to the similarity discovered between generic notions. We utilize the IDF to assess the specificity of a word, which is calculated as the sum of the documents in the documents divided by the sum of the documents containing that word. The edge weights are divided by the average IDF of the two edge nodes prior to running the EMM algorithm for each edge. For this tactic, we created the expression "EMM before." Word similarity and specificity are combined by the algorithm in a unique way. The similarity to the previously mentioned EMM is discovered using the scoring formula Sim (T1, T2) = (idfweight).

IV. RESULTS AND DISCUSSION

A. Data Set

In order to use a real-world dataset, we scraped the web for documents and metadata such as title, authors, and publication year.

TABLE I. DATA SET

Data set	No .of. documents
DBLP	100001
APS	450000
The Microsoft Research	5800 (pair of sentences)

B. Performance Evaluation

When calculating how similar two text documents are, the similarity based on content and citation takes into account both text- and link-based similarity. For the feature extraction document, an n-dimensional feature vector is used. Each feature vector's dimension correlates to a phrase in the text, and each dimension's value indicates the weight of the term to which it belongs. The phrase's weight reflects how authoritative and pertinent the document is on that topic compared to all the other documents in the dataset. From the DBLP dataset, the system's input documents are gathered. The documents with comparable features are taken into consideration for calculating similarity. The documents are composed of input text that conveys the contents. Two documents with several name aliases but the same author. Title, year, publishing information, number of papers, and size of all articles in 100001 format are among the fields. Total size is 5.23 MB.

The following formula provides the mean average precision metric to assess the efficacy of a similarity computation.

The formula for AP is (_(K=1)10P@K.Rel(k))/IRel (3)

Rel (k) is set to 1 if the document at position k is marked as relevant, where P @ k denotes the precision at position k. The query result set's size is indicated by the symbol |Res|.

C. Cosine Similarity

Utilizing the formula, the cosine similarity is utilized as the similarity metric to determine how similar the papers are.

|d1|= (d1[0]2+d1[1]2+d1[n]2) |d2|= (d2[0]2+d2[1]2+d2[n]2) (4)

where documents 1 and 2 are denoted by d1 and d2, respectively. In order to integrate the relevance scores of the phrases in the papers, the cosine similarity value between the documents is employed as the significance factor.

D. Dice

The dice similarity metric compares how similar the text documents are to one another. The quotient of similarity, or QS, is calculated using the formula QS = (2IXY)/(X+Y)(5), where X and Y are the numbers of elements in the two texts and QS is a number between 0 and 1.

E. BM25

It uses the BM25 similarity function. F (q, d) represents the term's frequency, which is determined by how many times the search term q appears in the document d. d. is the document d's word count (terms)

Result (q, d) = _(i=1)
^
q 〖idf(qi).(tf(q,d).(k1+1))/(tf(q,d)+k1.(1-b+b))〗 (6) (6)

The frequency of the word is expressed as the number of times the search phrase qi appears in the document d. The free parameters k1 and b correspond to the term's frequency.

F. Results

This section compares the accuracy of similarities computed using phrases from the title, abstract, and body to those computed using a variety of similarity metrics. We also compare their effectiveness to that of link-based, text-based, and hybrid methods. The success of

similarity computation is measured and evaluated using a variety of metrics, some of which include Cosine similarity, Dice, and BM25.

TABLE II. Experimental Results Comparison

Documents	D1 Relevance score	D2 Relevance score	D3 Relevance Score	D4 Relevance score
D1	1.0000	0.0889	0.0409	0.0985
D2	0.0884	0.9999	0.0319	0.1572
D3	0.0409	0.0319	1.0000	0.0793
D4	0.0985	0.1572	0.0793	0.9999

TABLE III. Similarity Score Comparison with Metrics

	Metric	D1 Similarity Score	D2 Similarity Score	D3 Similarity Score	D4 Similarity Score
Sim CC Approach	Cosine	0.54	0.32	0.21	0.48
	Dice	0.52	0.35	0.24	0.33
	BM25	0.55	0.22	0.31	0.29
Extended Maximum Matching Approach	Cosine	39.27	40.23	24.06	94.31
	Dice	1.83	1.23	1.06	1.38
	BM25	1.97	1.54	1.34	1.66

TABLE IV. RESULTS OF RANDOM GRAPH WALK

	Metric	Accuracy	F-Measure
Random Graph Walk Approach	Cosine	68.7	78.7
	Dice	70.8	80.1
	JS	68.8	80.5

V. CONCLUSION

This study analyzed works in three categories—knowledge-based approaches, corpus-based methods, and string-based methods—to determine the state of the art in semantic similarity methods at the time. 25 research that employ diverse methods and tactics to gauge semantic similarity have been thoroughly investigated and analyzed. The analysis demonstrates that knowledge-based and corpus-based methods are frequently employed to determine semantic similarity and that these methods produce encouraging outcomes.

REFERENCES

[1] S. Yoon, S. Kim, J. Kim (2017), 'On computing text-based similarity in scientific literature', In Proceedings of the 20th International Conference Companion on World Wide Web,Vol.34, pp. 169–170

[2] Abdul Ahad, Muhammad Fayaz, Abdul Salam Shah (2016), 'Navigation through Citation Network Based on Content Similarity Using Cosine Similarity Algorithm', International Journal of Database Theory and Application ,Vol.9, No.5, pp.9-20

[3] Seok-Ho Yoona, Sang-Wook Kima, Sunju Parkb (2016),'C-` BN

[4] Rank: A link-based similarity measure for scientific literature databases', Information Sciences, Vol.326,pp. 25–40

[5] Yoo Kyung Jeong,Min Song(2016), 'Applying Content-based Similarity Measure to Author Co-citation Analysis', Conference paper,Vol.32,pp.45-49

[6] Christian Paul, Achim Rettinger, Aditya Mogadala, Craig A. Knoblock, Pedro Szekely (2014), 'Efficient Graph-based Document Similarity', Information Sciences Institute, University of Southern California,Vol.91,pp.12-16.

[7] Rouzbeh Meymandpour,Joseph G.Davis (2016), 'A semantic similarity measure for linked data: An information content-based approach', Elsevier Transactions on Knowledge-Based Systems, Vol.109 , pp.276–293

[8] Majid Mohebbi and Alireza Talebpour (2016), 'Texts Semantic Similarity Detection Based Graph Approach', The International Arab Journal of Information Technology, Vol.13, No. 2, pp.213-217.

[9] Zia Ul-Qayyum and Wasif Altaf (2015), 'Paraphrase Identification using Semantic Heuristic Features/Research Journal of Applied Sciences', Engineering and Technology , Vol.56, pp.4894-4904.

[10] Justin J. Miller (2013), 'Graph Database Applications and Concepts with Neo4j' , In Proceedings of the Southern Association for Information Systems Conference, USA,Vol.72.pp.399-401.

[11] Majid Mohebbi, Alireza Talebpour (2014), 'Graph Based Measure of Text Semantic Similarity Using WordNet as a Knowledge Base', International Journal of Advanced Research in Computer Science & Technology, Vol.87,pp.45-49.

[12] Soumyajit Ganguly and Vikram Pudi (2013), 'Combining Graph and Text Information for Scientific Paper Representation', Vol.98, pp.45-52.

[13] Masoud Reyhani Hamedani, Sang-Wook Kim, JacSi (2013), ' An accurate and efficient link-based similarity measure in graphs', Information Sciences, Vol.414, pp.203–224

[14] Weiguo Zheng, Lei Zou, Yansong Feng, Lei Chen Dongyan Zhao(2013), ' Efficient SimRank-based Similarity Join Over Large Graphs', In Proceedings of the VLDB Endowment, Vol. 6, No. 7,pp.53-58.

[15]cSaschaRothe and HinrichSchutze (2014), CoSimRank: AFlexible&EfficientGraph-TheoreticSimilarityMeasure, In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Vol.54, pp.1392–1402.

[16] Nitesh Pradhan, Manasi Gyanchandani, Rajesh Wadhvani (2015), 'A Review on Text Similarity Technique used in IR and its Application', International Journal of Computer Applications ,Vol.120 – No.9, pp.561-566.

[17] Lingling Meng, Runqing Huang and Junzhong Gu (2016), A Review of Semantic Similarity Measures in WordNet, International Journal of Hybrid Information Technology Vol. 6, No. 1, pp.1827-1833.

[18] Maheshkumar B.Landge, Ramesh R.Naik (2015) , 'C. Namrata Mahender, Measuring Author Impression Using Cosine Similarity Algorithm', International Conference On Recent Advances in Computer Science,Vol.65,pp.1245-1249.

[19] Issa Atoum and Ahmed Otoom (2016), 'Efficient Hybrid Semantic Text Similarity using Wordnet and a Corpus ',International Journal of Advanced Computer Science and Applications, Vol. 7, No. 9,pp.52-57

[20] S. S. Sonawane and Dr. P. A. Kulkarni(2014), 'Graph based Representation and Analysis of Text Document: A Survey of Techniques', International Journal of Computer Applications, Vol.96, No.19.