International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 02 | Feb-2018 www.irjet.net p-ISSN: 2395-0072

AN EFFECTIVE DOCUMENT SIMILARITY APPROACH USING GRAMMATICAL LINKAGES IN SEMANTIC GRAPHS

Priya.V AP/CSE[1], K. Umamaheswai Prof/IT[2], Baskaran[3], Paarkavi [4]

[1,3,4] Dr.Mahalingam College of Engineering and Technology
[2] PSG college of Technology

Abstract - *Assessing the similarity of documents is at the core of many applications such as document retrieval and recommendation. Most similarity approaches operate on word-distribution based document representation-fast to compute, but problematic when documents differ in language, vocabulary or type, neglecting rich relational knowledge. Though, graph based approaches for document similarity leverage relational knowledge but it is infeasible in many applications, because of expensive graph operations. A semantic similarity approach is proposed for exploiting explicit hierarchical and traversal relations that generates a semantic graph and does the following process: expand the query document, store expanded document, pre search, apply topical intent approach and perform full search. This approach does not involve any grammatical linkages. So a new approach involving grammatical linkages using verbal intent technique is proposed and using this document similarity can be computed efficiently compared to other graph traversal approaches.*

1.INTRODUCTION

In traditional method the semantic similarities between the documents cannot be found accurately, because it only involves clustering algorithms which does not consider the semantic considerations. So the document can be easily modified and used for other purposes. To overcome this problem introducing the semantic meaning through Word Net has been widely used to find the similarities between the documents. Searching for related documents given a query document is a common task for applications in many domains. Established approaches measure text similarity statistically based on the distributional hypothesis, which states that words occurring in the same context tend to be similar in meaning. By inferring semantics from text without using explicit knowledge, word-level approaches become susceptible to problems arises when using distributional measures across heterogeneous documents, due to different vocabularies and text length or languages, each type may underlie a different word distribution, making them hard to compare. Here, it presents a scalable approach for related-document search using entity-based document similarity. A semantic similarity approach is proposed for exploiting explicit hierarchical and traversal relations that generates a semantic graph and does the following process: expand the query document, store expanded document, pre search, apply topical intent approach and perform full search. So a new approach involving grammatical linkages using verbal intent technique is proposed and using this document similarity can be computed efficiently compared to other graph traversal approaches.

The paper attempts to alleviate mentioned above problems its contribution can be summarized as follows.

1. We propose a modified similarity approach which uses the sematic similarities and verbal intent technique for comparing the documents and finding the similarities between the documents. Previous works have showed that exploiting the structural information can only improve the accuracy of similarity measurement, in this paper we explore that by using word net tool and verbal intent technique the similarity can be measured easily.

2. We introduce the semantic similarity which plays an important role natural language processing, informationretrieval text categorization. The semantic similarity method can be generally grouped into four types: path length based, information content based, feature based and hybrid measures. And verbal intent technique is used which generally matches the words based upon the topical intents.

3. We tell about how Pre-processing is done in the documents , and how semantics for the entities in the documents are identified and algorithms to find the entities which is main for finding the similarities between the documents, the algorithms to find the similarities is also discussed.

4. We also demonstrate the metrics for the calculating the similarities between the documents. The formula to find the similarities is also discussed. The paper also discusses about the some related works and presents similarity measured based on word net tool and the verbal intent technique. we have discussed about the existing system and we have identified the drawbacks in the existing system We have discussed about the earlier research of the system in which the comparison is done based upon the topical indents graphical similarities are viewed We have also discussed about the techniques which are used in the tokenization, entity identification, word net tool matching, knowledge graph generation and semantic similarity computation. Finally we conclude with results and the future enhancement work of the work.

2. LITERATURE SURVEY:

The authors Woo-Jong Ryu, Jung-Hyun Lee[1] defines that the existing semantic approaches to contextual advertising effectively match relevant ads to webpages in terms of topical intent. In this, it seeks to utilize the verbal intent, which complements topical intent, in semantic contextual advertising. Verbal intent describes what a user wants to do, i.e. action perspective, with topical intent. The results of performance evaluation on real-world datasets clearly show the efficacy of (verb, topic) associations. It generally seeks to identify related verbs for each topic on topical taxonomy.

The authors Christian Paul, AchimRettinger , Aditya Mogadala[2] states that document similarity can be calculated efficiently compared to other graph-traversal based approaches that experiments that a similarity measure that provides a significantly higher correlation with human notions of document similarity than comparable measures, this also holds for short documents with few annotations, document similarity can be calculated efficiently compared to other graph-traversal based approaches. During similarity calculation of the query document it expand query document, Store expanded document, pre search and full search is performed. Several approaches for graph based document were proposed based on the lexical knowledge graph word Net. This achieves a highly scalable solution by performing all knowledge graph related work in the Semantic Document Expansion pre-processing step.

GunesErkan, Dragomir R. Radev[3] describes about the stochastic graph-based method for computing relative importance of textual units and a detailed analysis of Lex rank approach and apply it to a larger data set. They discuss how random walks on sentence-based graphs can help in text summarization and also briefly discuss how similar techniques can be applied to other NLP tasks such as named entity classification, prepositional phrase attachment, and text classification. Graph-based centrality has several advantages over Centroid. First of all, it accounts for information subsumption among sentences. If the information content of a sentence subsumes another sentence in a cluster, it is naturally preferred to include the one that contains more information in the summary.It concludes by trying to make use of more of the information in the graph.

The authors KhuatThanh Tung, Nguyen Duc Hung, Le Thi My Hanh[4]describes, the similarity of two documents is gauged by using two string-based measures which are character-based and term-based algorithms.

In character-based method, n-gram is utilized to find fingerprint for fingerprint and winnowing algorithms, then Dice coefficient is used to match two fingerprints found. Intern-based measurement, cosine similarity algorithm is used. In this work, we would like to compare the effectiveness of algorithms used to measure the similarity between two documents It concludes that the selection of appropriate approaches and parameter settings will give better performance of similarity measurement between two documents.

Authors Ganggao Zhu, Carlos A.Iglesias[5] they provide a method for measuring the semantic imilarity between concepts in KGs such as Word Net and DBpedia.They proposed a semantic similarity method, namely wpath, to combine these two approaches, using IC to weight the shortest path length between concepts.

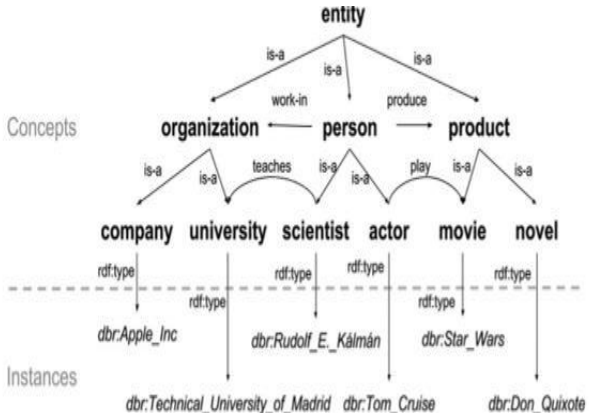


Figure 3.A tiny example of knowledge graph.

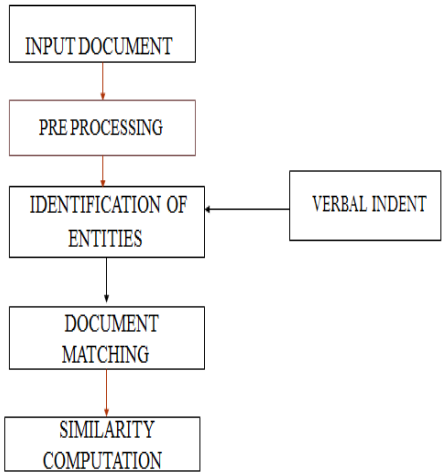
The authors Atulgupta, dharamveerkr.Yadav[6] proposes a metric for semantic relatedness calculation between pair of concepts which uses Tversky's feature based approach which takes into account the common and distinct feature of the two terms or concepts.

In the Word Net hierarchy, more specific and more Informative concept are there, where as when we move up in the hierarchy more Generalized and less Informative concepts are there. So depth of a concept in the Word Net hierarchy is a critical factor in similarity calculation. We take into consideration the depth of the specific concept in the Word Net hierarchy which is the deciding factor for determining the relevance of distinct feature specific to a concept in similarity calculation.

Above mentioned are the insights found in the former system, the following system overcomes the former system's drawback.

3. PROPOSED WORK:

In the proposed system, tokenisation is done at first level and then identification of entities is done with verbal indent technique by which document matching is done. Here, topical indent technique is applied and documents are kept classified according to various topics processed in graph databases. Finally, similarity is computed.



3.1 Architecture

The overall architecture of the proposed work is,

3.1.1.Tokenisation

1. Input the documents, does tokenization
2. Removes unwanted tags and special characters in the documents
3. Pre-processing is completed.

3.1.2.Entity Identification

1. Entities are identified in the document such as nouns, pronoun, adverb and verbs.
2. Thesemantic meaning for each word in the document is found using word net tool by using verbal indent technique.

3.1.3.Document Matching

By using semantics both the documents are compared and similarity is computed.

ALGORITHMS:

KNN ALGORITHM:

The KNN classifier algorithm is the k-nearest neighbour algorithm is the non-parametric method which is used for classifying parts of speech. In this algorithm the classification is made based on the closest training examples in the feature space. The output depends on whether knn is used for classification.

MAPPING ALGORITHM:

The mapping algorithm generally used for mapping of each of the phrases and finds the similar word which has the same meaning. Let L be the total number of documents which are given as the input, in which Ei is the first document for which the entities is i, let the function be the number of phrases in the sentence. If there are k phrases and they can be referred as j=1…k.

Consider another document Fi which contain another number of phrases which is the another document which contains some other phrases. By the algorithm, the similar words to which the original words are mapped and extracted from the word net tool and it is displayed.

Step1: Initially, after completion of tokenisation process it maps the words in the documents with those similar words in the word net tool.

Step 2: Then, fetches the similar words and displays

Step 3: Finally,mapping of original words in the document with similar words in word net tool is done.

4. COSINE SIMILARITY:

The cosine similarity method uses two documents as the inputs and in which the D1 and D2 are two documents in which it identifies the word and finds for the similar words in the documents by using the word net tool.Finally the documents the entities of the documents are identified and the similar words are identified from the word net tool and the document similarity using the semantic linkages is found. The total time to compute the similarity for each of the part of speech is also calculated.

Metrics for evaluation:

Document Score Computation is computed with matched (a1), a1 ∈ A1 denoting the annotation a2 ∈ A2 that a1 has an edge to in max Graph,

$$Sim_{doc(d1,d2)} = \frac{\sum_{a_{1i} \in A_1} (sim_{ent}(a_{1i} matched(a_{1i})))}{|A_1| + |A_2|}$$

each edge e = (v,w) carries siment(v,w) close to e's end towards v, and vice versa at its end towards w, for each annotation pair (a1,a2) compute entity similarity score.
d1 -document 1, d2 – document 2 (v,w)-edges,a1,a2-annotations, it refers documents.

5. EXPERIMENTAL RESULTS

Annotations,

A1(A2)=10,

A2(A1)=10

Sim(d1,d2)=(no of matched entities d1 + no of matched entities d2) /15

=(5+2)/15

=0.466

Similarity Score between documents is 0.466

6. CONCLUSIONS

Thus this method can be used for used for computing the similarity of the documents which can be mainly used in the educational institutions for computing the similar documents and to evaluate the marks. This system can be developed as an web application and the scanned documents can also can be given as the inputs and the similarities can be computed.

REFERENCES

[1] [1] A. Broder et al., "A Semantic Approach to Contextual Advertising," Proc. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, 2007, pp.559-566.

[2] [2] Christian Paul, AchimRettinger , Aditya Mogadala "Efficient Graph-based Document Similarity", International Sematic web Conference, the semantic web and latest advances,May 14,2016.

[3] [3] GunesErkan, Dragomir R. Radev, "LexRank: Graph-based Lexical Centrality as Salience in Text Summarization'', Journal of Artificial Intelligence Research,2004 (457-479)

[4] [4] KhuatThanh Tung, Nguyen Duc Hung, Le Thi My Hanh, "A comparison of algorithms used to measure the similarity between two documents", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 4 Issue 4, April 2015

[5] [5] Atulgupta, dharamveerkr.Yadav, "Semantic similarity measure using information content approach with depth for similarity calculation", International Journal of Scientific & Technology Research Volume 3 Issue 2, February 2014.

[6] [6] GanggaoZhu,CarlosA.Iglesias , "Computing Semantic Similarity of Concepts in knowledge Graphs", IEEE transactions on knowledge and data Engineering, vol.29,no .1,January 2017.

[7] [7] Philip Resnik , "Using Information Content to evaluate semantic similarity in a Taxonomy", ACM digital library 1995.

[8] [8] LinglingMeng,Runqing Huang, "Review of Semantic Similarity Measures in Wordnet", International Journal of Hybrid Information Technology, Vol 6,No.1,January 2013.

[9] [9] Julian Kupiec, "An Algorithm for Finding Noun Phrase Correspondences in Bilingual Corpora" ,Palo Alto.

[10] [10] Takagi, N., Tomohiro., M.: Wsl: Sentence similarity using semantic distance between words. In: SemEval. Association for Computational Linguistics (2015)

[11] [11] Nunes, B.P., Kawase, R., Fetahu, B., Dietze, S., Casanova, M.A., Maynard, D:Interlinking documents based on semantic graphs. Procedia Computer Science 22, 231–240 (2013)

[12] [12]Dataset at https://webfiles.uci.edu/mdlee/LeePincombeWelsh.zip