Chapter 13
Document Similarity Approach Using Grammatical Linkages with Graph Databases

V. Priya and K. Umamaheswari

13.1 Introduction

Document similarity assessment is helpful in exploring linked documents based on a source text. It is a useful mechanism for many fields dependent on text processing. Most systems measure text similarity [1] depending on the word distribution statistics. They measure the similarity assuming that the words have analogous meaning when they appear in the identical environment. These word-based techniques become vulnerable to many complications when they deduce some inference from text without using clear knowledge. When utilizing distributional measures, word-based approaches become inefficient for comparisons. Some of the reasons include document heterogeneity, varied vocabulary, text length, and languages.

In traditional way the semantic similarities between the documents cannot be found accurately because the semantic considerations are not used. So the document can be easily modified and used for other purposes. To overcome this problem introducing the semantic meaning from Word Net has been widely used to find the similarities between the documents. Conventional approaches use hypothesis based on text distribution statistics. They assume that two words might have same meaning when found in the similar situation. When documents are heterogenic in nature, concluding interpretations in the text with no clear knowledge might rise to have problems.

V. Priya (✉)
Dr. Mahalingam College of Engineering and Technology, Pollachi, Tamil Nadu, India

K. Umamaheswari
PSG College of Technology, Coimbatore, Tamil Nadu, India

A. Haldorai et al. (eds.), *EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing*, EAI/Springer Innovations in Communication and Computing, https://doi.org/10.1007/978-3-030-19562-5_13

Word distributional metrics are not feasible since documents with different vocabularies and length follow a diverse distribution of words. It becomes complex to compare these textual units. Conventional semantic similarity approaches utilizing semantic graphs are infeasible because of expensive operations. So a new approach involving grammatical linkages using verbal intent technique is proposed and using this document similarity can be computed efficiently in comparison with other state-of-the-art graph-based mechanisms.

Section 13.2 of the chapter details several techniques used for text similarity using graph-based approaches. Section 13.3 presents the work on the verbal intent-based document similarity identifier. Finally, Section 13.4 presents conclusion.

13.2 Literature Survey

This section presents a study on various techniques used in detection and computation of document similarity. Researches explored a variety of measures depending on textual units, graphical units, and semantic units.

Some of the commonly used text-based techniques in document similarity identification and computation are text-based and semantic-relation-based approaches. The authors [2] have utilized two measures which rely on character and term-based algorithms for computing the similarity of two documents. In the first method, n-gram is utilized to identify fingerprint using winnowing algorithms. Then Dice coefficient is adopted to find similarity in the two fingerprints identified. They have employed an algorithm to link noun expressions using an affiliated multi-lingual corpus.

Christian et al. [3] have analyzed that document similarity could be computed effectively compared to graphical unit-based methods by using similarity measure. The measure should provide a significantly higher connection with human notions of document similarity. The authors in [4] have employed random graphical approach for computing comparative prominence of textual units and a detailed analysis of Lex rank mechanism and applied it to a huge data set. They deliberate the helpfulness of applying random walks to sentence-based graphs would improve in text summarization. They also briefly explain the possibility of deploying such methods in NLP tasks such as classifying named entities, attaching prepositional phrases, and text categorization. Graph-based centrality has quite a few advantages over Centroid method. In [5] document similarity has been studied using semantic similarity. Semantic similarity [6–10] and their measures had been explored in literature extensively. In these schemes semantic similarity among the concepts found in Knowledge Graphs such as WordNet and DBpedia are measured using wpath metric. This combines information content to identify the length of the shortest path between any two concepts.

13.3 Proposed Work

This section presents the work on text similarity detection using verbal intent and graph databases. The design of the proposed system is shown in Fig. 13.1. This consists of a Data preprocessing module which tokenizes and tags the entities from the sentences. Tokenization is done with Part-of-Speech (POS) tagging. Identification of entities is done with POS tags [11]. They are used for generating knowledge graph with links. Now weights are assigned using verbal intent technique along with graph database generation. Lastly, similarity is calculated based on the weights from the verbal intents and links in the knowledge graph.

13.3.1 Architecture

The complete design of the system is shown (in Fig. 13.1).

13.3.1.1 Tokenization

In the given input documents, tokenization is done. Tokenization is performed by splitting the sequence of strings. This splitting results in individual elements called as tokens which are keywords, phrases, symbols, and other special elements. Spaces, tags, and special characters which are considered to be irrelevant are removed.

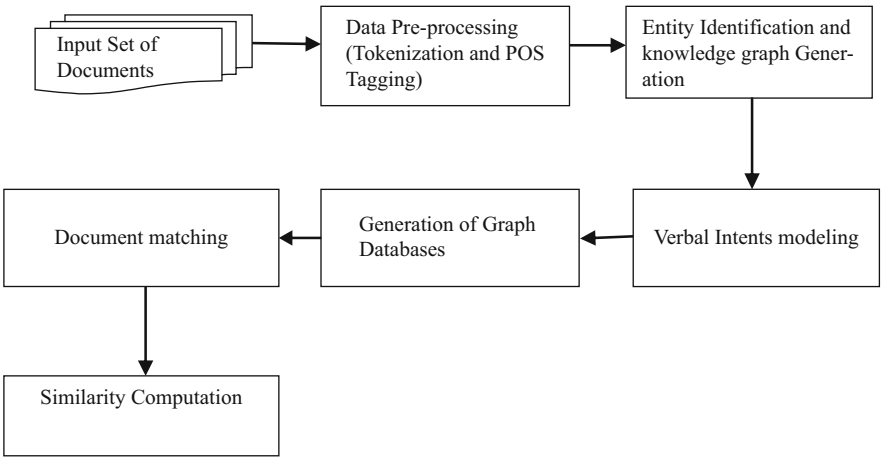

Fig. 13.1 System architecture

13.3.1.2 Verbal Intent Modelling

The verbal intent of any document is computed as follows. A verbal intent describes the author's action perspective, based on the subject of the document. Given document and its noun terms, a verbal intent is modelled as follows: First verbal feature is defined and represented as a vector of weighted values as shown in Eq. (13.1):

$$V_n = W_s + W_a \tag{13.1}$$

where every facet of a vector of verb V_n matches conventional relevant verbs intended for the noun associated, W_s and W_a represents the weight from synonyms and adverbs of the referenced document, respectively. Given a document, computation of verbal weights is performed using cosine similarity between each term vector of document and d_n is done as given in Eq. (13.2).

$$W_n = \text{cosine}\,(V_n, d_n) \tag{13.2}$$

where the weight of every term in the term vector is calculated using the usual tf-idf scheme adopted from the Vector Space prototype [7] of the document.

The partial results are obtained with cosine similarity metric with summarized data. It is likely that the performance improvement could be observed using semantic knowledge bases and graph databases.

13.4 Conclusion

Document similarity systems are found to have enormous usage for many applications like plagiarism detection, template matching, and so on. The approach using verbal intents for document similarity computation was deliberated. The system is expected to generate feasible results for small documents such as short summaries as well as large size documents in the corpus. Further the system could be improved in introducing intents for all entities to improve semantic similarity. Other promising future directions include extending the system for online documents and template verification in IT contract services which can improve customer relationship.

References

1. R. Anna, Z. Silvia, Assessing semantic similarity of texts—methods and algorithms, in *Proceedings of the 43rd International Conference Applications of Mathematics in Engineering and Economics*, 2010, pp. 1–8

2. K. Julian, An algorithm for finding noun phrase correspondences in bilingual corpora, in *Proceedings of the 31st Annual Meeting on Association of Computational Linguistics*, 2012, pp. 17–22
3. P. Christian, R. Achim, M. Aditya, Efficient graph-based document similarity, in *Proceedings of the 13th International Conference on the Semantic Web. Latest Advances and New Domains*, vol 9678, 2016, pp. 334–349
4. E. Gunes, R. Dragomir, LexRank: graph-based lexical centrality as salience in text summarization. J. Artif. Intell. Res., 457–479 (2015)
5. Z. Ganggao, A. Carlos, Computing semantic similarity of concepts in knowledge graphs. IEEE Trans. Knowl. Data Eng. **29**(1), 72–85 (2017)
6. R. Philip, Using information content to evaluate semantic similarity in a taxonomy, ACM Digital Library, 1995, pp. 448–453
7. E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Paşca, A. Soroa, A study on similarity and relatedness using distributional and WordNet-based approaches, in *Proceedings of Human Language Technology Annual Conference North American Chapter Association of Computational Linguistics*, 2009, pp. 19–27
8. A. Broder et al., A semantic approach to contextual advertising, in *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2007, pp. 559–566
9. J.-H. Lee et al., Semantic contextual advertising based on the open directory project. ACM Trans. Web **7**(4), 1–24 (2013)
10. N. Takagi, M. Tomohiro, Wsl: sentence similarity using semantic distance between words, in *Proceedings of the Ninth International Workshop on Semantic Evaluation*, 2015, pp. 128–131
11. A. Gupta, D.K. Yadav, Semantic similarity measure using information content approach with depth for similarity calculation. Int. J. Sci. Technol. Res. **3**(2), 165–169 (2014)