ISSN:0975-9646

Plagiarism Detection Considering Frequent Senses Using Graph Based Research Document Clustering

P.Kalyan Chakravarthy[1], J.Bindu Kavya[2], K.Sireesha[3], D.Mounika[4], B.Sruthi[5]

[1] Assistant Professor, [2,3,4,5]Student, Department of CSE,
Lendi Institute Of Engineering & Technology, Vizianagaram ,AP, India

AbstractA new, graph based research document clustering technique (GRD-Clust) is introduced based on frequent senses rather than frequent keywords as per the traditional document clustering techniques.GRDClust presents text documents as hierarchal document-graphs and utilizes an Apriori paradigm to find the frequent sub graphs, which reflect frequent senses based on support and confidence. We highlight the different types of plagiarism and address the issues of plagiarism of text, plagiarism of ideas, mosaic plagiarism, self-plagiarism, and duplicate publication. Different documents eschewed of plagiarism by identifying the alleged terms are considered. An act of plagiarism can have several repercussions when an article does not score high on clarity or lacks conciseness, the deficiency is typically unintentional.

1. INTRODUCTION

Plagiarism is the wrongful presentation of somebody else's work or idea as one's own without adequately attributing it to the source. Plagiarizers use or take intellectual property without permission or giving credit such as words by rearranging, other's ideas, processes, and result. Our approach is motivated by typical human behaviour, when given a task of organizing multiple documents. As an example, consider the behaviour of editor, who needs to organize multiple research papers into a single book volume, with a hierarchical table of contents. Typically, research papers, even when coming from the same area, are written in multiple writing styles, on different levels of detail, and in reference to different aspects of an analyzed area. Instead of searching for identical words and counting their occurrences, like many well-known computer-based text clustering techniques do [2]–[4], the human brain usually remembers only a few crucial keywords representing senses, which provide the editor with a compressed representation of the documents. These senses are then used to fit a given research paper into a book organization scheme, reflected by the table of contents. In our work, we replace editor's knowledge with ontology and use it to discover common senses that can then be used to organize documents.

In GRDClust, we construct document-graphs from text documents and apply an Apriori paradigm [18] for discovering frequent sub graphs from them. We utilize a hierarchic representation of English terms, Word Net [1], to construct document-graphs. Since each document can be represented as graphs of related terms, they can be searched for frequent sub graphs using graph mining algorithms. We aim to cluster documents depending on the similarity of the sub graphs in the document-graphs. GRDClust enables clustering of documents providing humanlike sense-based searching capabilities, rather than focusing only on the co-occurrence of frequent terms. It follows the way human beings process the text data. As the outcome of GRDClust, we achieve sub graphs of meaningful senses.

2. LITERATURE REVIEW

Plagiarism of text is also called "word-for-word "plagiarism. "…copying a portion of text from another source without giving credit to its author and without enclosing the borrowed text in quotation marks."Earlier, plagiarizing text from an article also required considerable hard work. One had to visit libraries and go through volumes of literature and read several textbooks to be able to copy relevant ideas and text. Even access to such resources was limited. Today, with the advancement of technology, plagiarism is easy. The practice seems to have increased manifold due to readily available internet access, simply because information is easily available online which can then be copied."Cut-copy-paste "seems to be happening across the world and is significantly prevalent in India as well. We have to understand that though technology makes plagiarism easy, it also makes detection of plagiarism even easier. Ethical medical writers must always acknowledge the original source of the idea, text, or illustration.

The writer must read the instructions to authors to know what style they need to use There are both paid and free online software that can easily detect even short phrases that are copied verbatim from the original source we develop a graph-based document technique for clustering text documents. We represent the documents of a repository as graphs. The benefit of GRDClust is that it is able to group documents in the same cluster even if they do not contain common keywords, but still possess the same sense. Existing clustering techniques cannot perform this sort of discovery [2]–[4] or do this work only to a limited degree

Our system depends on background knowledge of the English language ontology that is constructed. We aim to cluster documents depending on the similarity of the sub graphs in the document-graphs. GRDClust enables clustering of documents providing humanlike sense-based searching Capabilities, rather than focusing only on the co-occurrence of frequent terms. It follows the way human beings process the text data. As the outcome of GRDClust, we achieve sub graphs of meaningful senses. The current

methods of plagiarism detection relay on the comparison of small text such as character, n-gram, chunk or terms. Suppose we have a document content ten sentences for which the graph to be generated. The consideration of small text unit for detecting of similarity between the original document graph and suspected document graph lead to huge number of comparison GRDClust offers a fully automated system that utilizes Apriori-based sub graph discovery technique to harness the capability of sense-based document clustering.

3. SYSTEM OVERVIEW

This section portrays the techniques used for sense discovery and document clustering in GRDClust.

3.1. *Document-graph construction algorithm*

GRDClust utilizes BOW Toolkit [6] and WordNet 2.1 taxonomy to convert a document to its corresponding document-graph (Table 1). We utilized the WordNet's noun taxonomy, which provides a *hypernymy-hyponymy* relation between concepts and allows constructing a Concept Tree with up to 18 levels of abstractions. A *concept* is a set of synonymous words named *synset*. All nouns in WordNet are merged to a single topmost synset (i.e.,*{entity}*).

TABLE 1. ALGORITHM FOR CONSTRUCTION OF DOCUMENT-GRAPHS

(1) For each document *Di*, construct a document-graph *Gi*, where $1<i<n$, and *n* is the total number of documents {
(2) For each keyword, *kj* where $1<j<m$ and *m* is the number of keywords in document *Di* {
(3) Traverse WordNet taxonomy up to the topmost level. During the traversal, consider each synset as a vertex. *E* is considered as a directed edge between two vertices *V1* and *V2*, iff *V2* is the hypernym of *V1*.
(4) *E* is labeled by *V1::V2*. If there is any repeated vertex or edge that was detected earlier for another keyword *kt* $(t \neq j)$ of the same document, *Di*, do not add the repeated vertices and edges to *Gi*, otherwise, add vertices and edges to *Gi*.
(5) } *// End of "For each keyword"*
(6) } *// End of "For each document"*

3.2 *Clustering text documents*

Agglomerative clustering to group documents together. We construct dissimilarity matrix for every pair of document-graphs. The widespread use of on-line publishing of text promotes storage of multiple versions of documents and mirroring of documents in multiple locations, and greatly simplifies the task of plagiarizing the work of others. We evaluate two families of methods for searching a collection to find documents that are co derivative, that is, are versions or plagiarisms of each other. The first, the ranking family, uses information retrieval techniques; extending this family, we propose the identity measure, which is specifically designed for identification of co derivative documents. The second, the fingerprinting family, uses hashing to generate a compact document description, which can then be compared to the fingerprints of the documents in the collection. We introduce a new method for evaluating the effectiveness of these techniques, and demonstrate it in practice.

Using experiments on two collections, we demonstrate that the identity measure and the best fingerprinting technique are both able to accurately identify co derivative documents. However, for fingerprinting parameters must be carefully chosen and even so the identity measure is clearly superior. Document clustering techniques mostly rely on single term analysis of the document data set, such as the vector space model. To achieve more accurate document clustering, more informative features including phrases and their weights are particularly important in such scenarios. Document clustering is particularly useful in many applications such as automatic categorization of documents, grouping search engine results, building taxonomy of documents, and others. This article presents two key parts of successful document clustering. The first part is a novel phrase-based document index model, the document index graph, which allows for incremental construction of a phrase-based index of the document set with an emphasis on efficiency, rather than relying on single-term indexes only.

It provides efficient phrase matching that is used to judge the similarity between documents. The model is flexible in that it could revert to a compact representation of the vector space model if we choose not to index phrases. The second part is an incremental document clustering algorithm based on maximizing the tightness of clusters by carefully watching the pair-wise document similarity distribution inside clusters. The combination of these two components creates an underlying model for robust and accurate document similarity calculation that leads to much improved results in Web document clustering over traditional methods.

4.CONCLUSION:

GRDClust presents a new technique for clustering text documents based on co-occurrence of frequent senses in the documents for plagiarism check. The developed novel approach offers an interesting, sense-based alternative to the commonly used technique for text documents and detecting the plagiarism. Unlike traditional systems, GRDClust harnesses its clustering capability from the frequent senses discovered in the documents. It utilizes graph-based mining technology to discover frequent senses. GRDClust is an automated system and minimal user interaction is required as plagiarism detects similarities in different documents. In the close future, we want to look carefully at the concept of the inexact matching of sub graphs [7], as we believe it can be used effectively during our clustering process. We expect that the inexact matching would allow us to select only larger sub graphs generated by the Apriori approach, which could further decrease computational costs involved in the phase of frequent sub graph candidate analysis.

ACKNOWLEDGEMENT:

We convey our deep sense of gratitude to Lendi Institute of Engineering and Technology We express our sincere thanks to our beloved Principal Dr.V.V.Rama Reddy, Head of the Department and Mr.A.Rama Rao ,Mr.P.Kalyan chakravarthy and Mr.Kartheek for their valuable guidance in presenting this paper

REFERENCES:

[1] Miller G.A. and Charles W.G., "Contextual Correlates of Semantic Similarity", *Language and Cognitive Processes*, vol. 6(1), 1991, pp. 1–28.

[2] F. Sebastiani, "Machine learning in automated text categorization", *ACM Comp. Surveys*, vol. 34(1), 2002, pp. 1–47.

[3] C. D. Manning and H. Schutze, "Foundations of Natural Language Processing", *MIT Press*, 1999.

[4] C. Cleverdon, "Optimizing convenient online access to bibliographic databases", *Inf. Survey and Use*, vol. 4(1), 1984, pp. 37–47.

[5] S. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman, "Indexing by latent semantic analysis", *Journal of the Society for Inf. Science*, vol. 41(6), 1990, pp. 391–407.

[6] A. McCallum, "Bow: A Toolkit for Statistical Language Modeling, Text Retrieval, Classification and Clustering", *http://www.cs.cmu.edu/~mccallum/bow/*

[7] N. Ketkar, L. Holder, D. Cook, R. Shah and J. Coble, "Subdue: Compression-based Frequent Pattern Discovery in Graph Data", *Proc. of the ACM KDD Workshop on Open- Source Data Mining,* August 2005, pp. 71–76.

AUTHORS

P.KalyanChakravarthy is now working as Assitant Professor in Lendi Institute of Engineering and Technology.He has been training students in datamining and Network Security and presented various papers in these fields .

J.Bindu Kavya pursuing B.Tech (C.S.E) in Lendi Institute of Engineering and Technology.Her interests are data mining, software development and doin things innovatively.

K. Sireesha pursuing B.Tech(C.S.E) in Lendi Institute of Engineering and Technology.Her interest is data mining.

D.Mounika pursuing B.Tech(C.S.E) in Lendi Institute of Engineering and Technology.Her interest is data mining.

B.Sruthi pursuing B.Tech(C.S.E) in Lendi Institute of Engineering and Technology.Her interest is data mining.