EFFECTIVE INTRINSIC PLAGIARISM DETECTION USING CLUSTERING METHOD

S.Prasanth

PG Scholar

Sri Ramakrishna Engineering College

Coimbatore

iamprasanthsrec@gmail.com

Mr. B.Saravana Balaji

Assistant Professor (Sr.G)

Sri Ramakrishna Engineering College

Coimbatore

saravanabalaji@ymail.com

Abstract

The contribution of this work relates to the modeling of writing style. We use a model for writing style quantification, for finding significant deviations in a document's writing style, these differing segments could have been plagiarized, and are probably useful as a starting point to search for possible source candidates. But in this work, one important issue is if more than one author had written the document then the existing method will indicate as the plagiarized content. To overcome this problem we introduce a clustering method. In this technique we first Select the 50 most frequent words from the file and Determine frequency by paragraph for these 50 words. Extract the nouns from the 50 most frequent words (excluding stop words). For the highest frequency noun, create a cluster and remove from consideration all other nouns enclosed by this – i.e. occurring in the same paragraphs. Repeat this step to produce new clusters from the remaining nouns. Based on this clustering approach we can obtain the accuracy result in the plagiarism detection. Experimental results is shown that the proposed system is very effective than the existing system.

1.INTRODUCTION OF PROJECT

Plagiarism defined as "unacknowledged copying of documents or programs". The documents are copied from the various sources. Analyses were undertaken by the academic community to deal with student plagiarism. With the growth of content found in the Web, people can find nearly

everything they need for their written work, but detection of such documents can become a difficult task. In order to discriminate plagiarized documents from non-plagiarized documents, a correct selection of text features is a key aspect. There are many types of plagiarism such as copy and paste, redrafting of the text, copying of idea, and plagiarism through translation from one language to another. Nowadays, many documents are available on the internet and are easy to access. Due to this availability, users can easily create a new document by copying and pasting from these resources. Sometimes users can reword the plagiarized part by replacing the word with their synonyms. This kind of plagiarism is difficult to be detected by the traditional plagiarism detection system.

Generally speaking, the task of plagiarism detection from an algorithmic point of view can be divided into two main strategies those that utilize only information within the suspected document, denominated intrinsic plagiarism detection, and it compares the suspected document against a set of possible sources (ideally, but unrealistically, the entire Web). Intrinsic plagiarism detection [1] tends to discover plagiarism by analyzing only the suspicious document, to identify the segments that are written by another person. Current algorithms usually use writing style modeling technique that used for searching meaningful variations. External plagiarism detection refers to the task of comparing the suspected document against possible sources, But in Intrinsic plagiarism we use a model for writing style that tends to find significant deviations in a document's writing style these differing segments could may get plagiarized, and are probably useful as a starting point to search for possible source candidates.

2. MOTIVATION

Many documents are available on the internet and are easy to access. Due to this availability, users can easily create a new document by copying and pasting from these resources.

Sometimes users can reword the plagiarized part by replacing the word with their synonyms. Motivation of the paper is to find the most plagiarism content that should be copied from anywhere identified in the efficient manner. Further it helps to as plagiarism detection process[2] in applications to user or individual publish their journals .

3. OBJECTIVE

Most empirical studies and analysis were undertaken by the academic community to deal with student plagiarism. In order to discriminate plagiarized documents from non-plagiarized documents, a correct selection of text features is a key aspect. The main objective of the paper is to find the more accurate plagiarism content in the documents with similar meaning and concepts are correctly identified in the efficient manner.

4. PROBLEM DEFINITION

The current plagiarism detection system was found to be too slow and takes too much time for checking. The matching algorithms are also dependent on the text' s lexical structure rather than semantic structure. Therefore, it becomes difficult to detect the text paraphrased semantically. The big challenge is to provide plagiarism checking with appropriate algorithm in order to improve the percentage of finding result and time checking. The important question for the plagiarism detection problem in this study is whether it is possible to apply new techniques such as Semantic Role Labeling to handle plagiarism problems for text documents.

5. SYSTEM ANALYSIS

5.1 EXISTING SYSTEM

We do text mining, exploring the use of words as a linguistic feature for analyzing a document by modeling the writing style. The main goal is to discover deviations in the style, searching for segments of the document that could have been written by another person. This can be considered as classification problem using self-based information, outliers are the paragraphs with significant deviations in style which called intrinsic plagiarism detection approach does not relies only on the use of words, so it is not language specific. In the following, some of the core ideas developed in this research is presented:

- To be able to distinguish different authors within the same document, one must characterize the writing style present in the text.
- The use of ''n-gram profiles'' [3] compares segments of the document against the whole document. This approach works based on the assumption that the document has a main author, who wrote the majority, if not all, of the text. Therefore, it is logical that the comparison between

the style of a particular segment with the whole document style could lead to detections of important variations, meaning that other authors are involved.

- Based on reading and contemplation, one of the characteristics that was shown to be of interest is the author's use of words. Different authors tend to use different words to write their ideas, whether on the same topic or not.

These ideas lead to the following intuition for the development of the algorithm: If some of the words used in the document are author-specific, one can think that those words could be concentrated in the paragraphs (or more generally, in the segments) that the mentioned author wrote.

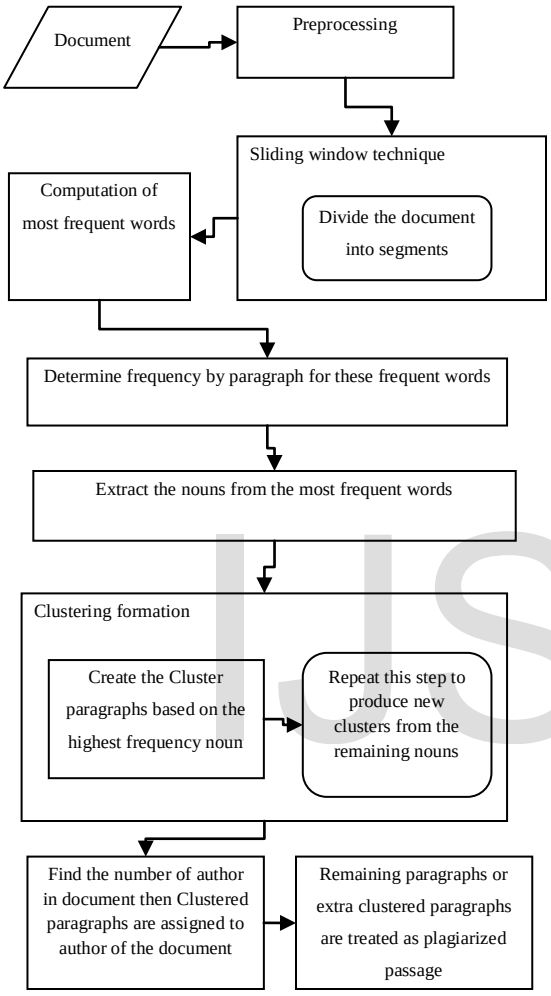
5.2 PROPOSED SYSTEM

To overcome the problem of existing system, we proposed a technique such as clustering method. In this proposed work, first the given document is preprocessed then divides the document into segments by using sliding window technique. Sliding window technique is used for divide the document by given length of the sliding window. For example one sliding window length has 400 words. After dividing the document, we extract the frequent words in the given document. For those frequent words, we determine the paragraphs based on the frequent words occurred in that paragraph.

Similarity function is used to find the distance of the frequent words with the paragraphs. Then extract the nouns from the most frequent words. For the highest frequency noun, create a cluster. By highest frequency nouns we clustered the paragraphs based on that. Next we create a new cluster of paragraphs for next priority frequent nouns.

If some paragraphs are not clustered then it means that paragraphs are plagiarized content or passage. After clustering formation we compare the number of cluster created with the number of author in the given document. If the number of cluster formation is higher than the number of author then the extra paragraph cluster is treated as coping content. From this clustering approach[4] we can obtain the more accuracy plagiarism detection result than the existing system.

6. ARCHITECTURE DIAGRAM FOR PROPOSED SYSTEM



7. SYSTEM IMPLEMENTATION

7.1 Preprocessing

Data pre-processing is an important step in the data mining process. The phrase "garbage in, garbage out" is particularly applicable to machine learning projects and data mining. Data-gathering methods[6] are loosely controlled, which results in out-of-range values (e.g., Income: −100), impossible data combinations,missing values, etc. Analyzing data has not carefully screened for problems that can produce misleading results. If there is much irrelevant and false information present, then knowledge discovery during the training phase is more difficult. Data preparation[5] and filtering steps can take considerable amount of processing time. Data pre-processing includes cleaning, normalization, transformation, feature extraction and selection, etc. First, the document is preprocessed by removing numbers and all other characters that do not belong to the a–z group. All characters are considered lowercase.

7.2 Sliding window technique

The sliding window parameters operate on letter characters. That is, a window length of *l* characters means that the window should contain *l* letter characters. Note that all the other characters (digits, spaces, punctuation, etc.) are not removed. Therefore, if *l*=1,000, a window may contain 1,200 characters (this is the real window length) in total from which 1,000 are letter characters. This procedure assures that all the text windows

will have the same number of letter (or content) characters and the formatting of the text will not significantly affect the style change function. The complete document is clustered creating groups C. As a first approach, these groups or segments c ∈ C

are created using a sliding window of length m over the complete document. Afterwards, for each segment c ∈ C, a new frequency

vector v_c is computed, which is used in further steps to compare whether a segment deviates with respect to the footprint of the complete document.

7.3 Intrinsic plagiarism evaluation

The general footprint[6] or style of the document is represented by the average of all differences computed for each segment and the complete document. Note that every segment is compared against the whole document only in terms of the words present in the segment. Also, this algorithm considers that if certain words are only used in a certain segment, the comparison of that segment against the whole document would lead to a low value, because the frequency of those words would be the same in both the whole document and in the segment. At last all segments are classified according to their distance with respect to the document's style. The main function in Algorithm 1, fifth line[7] , computes the differences in the use of words of two segments. The function is constructed so segments of the document that have many words that are exclusively in that segment will have a low value. This idea is generated based upon the use of words present be stable, with at least a high proportion of the words used throughout the document. Since the algorithm considers the information of each document to construct and evaluate variations in style, the function remains somewhat stable over varying document lengths[8]. The strong assumption here is that the majority of the text was written with the same writing style; otherwise no reliable information could be extracted from this model.

7.4 Semi- supervised clustering

In the training phase, we have given the training datasets with the clusters. So in the testing phase we have given the testing datas to find the clusters. Based on the training dataset we are giving the constraints for the testing dataset[9] . Based on the constraints the test data are clustered. The constraints are such as must link and cannot link. Then extract the nouns from the most frequent words. For the highest frequency noun,

create a cluster. By highest frequency nouns we clustered the paragraphs based on that. Next we create a new cluster of paragraphs for next priority frequent nouns[10] . If some paragraphs are not clustered then it means that paragraphs are plagiarized content or passage. After clustering formation we compare the number of cluster created with the number of author in the given document. If the number of cluster formation is higher than the number of author then the extra paragraph cluster is treated as coping content.

8. RESULTS AND DISCUSSION

9. CONCLUSION

In this study we explore the problem of text plagiarism and the possibility of its detection by the use of computer algorithms. With the rising utilization of digital documents and the Web, plagiarism is increasing as well. In view of this, many approaches to detect digital plagiarism have been introduced, and as seen, huge progress is being made in the field of automatic plagiarism detection. One of the first problems the systems face is the collection of possible sources to compare the suspected documents with. It is common that the ideal and real sources are not always available.Considering the latter issue, algorithms that do not rely on the available sources are being studied. Hence it is called intrinsic plagiarism detection concept was introduced. The idea, to analyze the document looking for variations that could hint at plagiarized passages, was recently tested and studies utilizing different writing style markers are being introduced. We study a self-based information algorithm, whose basic idea is the use of a function to quantify the writing style based solely on the use of words. But in this work, one important issue is if more than one author was written the document then the existing method will indicate as the plagiarized content. To overcome this problem we introduce a clustering method. Based on this clustering approach we can obtain the accuracy result in the plagiarism detection

REFERENCE

1. Baayen, H., van Halteren, H., & Tweedie, F. (1996). Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. Literary and Linguistic Computing, 11, 121–132.
2. Bao, J.-P., Shen, J.-Y., Liu, X.-D., Liu, H.-Y., & Zhang, X.-D. (2004). Semantic sequence kin: A method of

document copy detection. In H. Dai, R. Srikant, & C. Zhang (Eds.), Advances in knowledge discovery and data mining. Lecture notes in computer science (Vol. 3056, pp. 529–538). Berlin/Heidelberg: Springer.

3. Barrón-Cedeño, A., Basile, C., Degli Esposti, M., & Rosso, P. (2010). Word length ngrams for text re-use detection. In A. Gelbukh (Ed.). Computational linguistics and intelligent text processing. lecture notes in computer science (Vol. 6008, pp. 687–699). Berlin/Heidelberg: Springer.
4. Berry, M. W., Dumais, S. T., & O'Brien, G. W. (1995). Using linear algebra for intelligent information retrieval. SIAM Review, 37, 573–595.
5. Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In Proceedings of the fifth annual workshop on computational learning theory COLT '92 (pp. 144–152). New York, NY, USA: ACM.
6. Bravo-Marquez, F., L'Huillier, G., Ríos, S. A., & Velásquez, J. D. (2011). A text similarity meta-search engine based on document fingerprints and search results records. Proceedings of the 2011 IEEE/WIC/ACM international conferences on web intelligence and intelligent agent technology – WI-IAT '11 (Vol. 01, pp. 146–153). Washington, DC, USA: IEEE Computer Society.
7. Ceska, Z. (2008). Plagiarism detection based on singular value decomposition. In GoTAL '08: Proceedings of the sixth international conference on advances in natural language processing (pp. 108–119). Berlin/Heidelberg: Springer.
8. Chow, T. W. S., & Rahman, M. K. M. (2009). Multilayer SOM with tree-structured data for efficient document retrieval and plagiarism detection. Transactions on Neural Networks, 20, 1385–1402.
9. Grieve, J. (2007). Quantitative authorship attribution: An evaluation of techniques. Literary and Linguistic Computing, 22, 251–270.
10. Grman, J., & Ravas, R. (2011). Improved implementation for finding text similarities in large sets of data – notebook for pan at CLEF 2011. In V. Petras, P. Forner, & P.