IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

Intrinsic Plagiarism Detection for Text Based Features Pattern

P.Rubini*[1], Ms. S.Leela[2]

*[1]Post-Graduate Student, [2]Assistant professor, Department of Computer Science and Engineering,
Karunya University, India
rubijesus1@gmail.com

Abstract

Plagiarism detection means detecting the document whether copied or stealing from the other document. The main goal is to detect the word by analyzing the writing style using technique intrinsic plagiarism detection. Text mining is used to extract the useful information from the text. Intrinsic plagiarism detection is used to take the few words from the document and then it compared to the original document whether it's plagiarized or not. In addition, it performs the modern method in intrinsic plagiarism detection such as Recall, Precision, F-measure and Granularity. stylometric frequent pattern of authors can be extracted and mined along with lexical character features, lexical word features and syntactical features. The proposed work detects the plagiarism for documents based on collecting author profiles of documents.

Keywords: Intrinsic plagiarism detection, Plagiarism detection, copied document, style modeling.

Introduction

Plagiarism is occurring in everyday topics, for example: in research, academics, empirical studies, literature, etc. To reduce this plagiarism detection educate the people not do it, encourage to think in his way and help people not to do the mistake. These prevention method is not reduce, thus the solution is not trivial. The different kinds of plagiarism is,

1. Exact copy: it copy and paste the entire document without any change in the document.

2. Idea: it copies only their idea and the content are different from others.

3. Paraphrasing: it copies the entire document with small changes in the content by giving in active and passive words.

Text mining is used to extract the useful information from the text. Those texts are used to collect in a cluster format. Each and every segment is in a form of cluster. This segment will detect the deviations in the writing style.

Related Works

Intrinsic Plagiarism Detection

Intrinsic plagiarism detection is used to refer the intrinsic algorithm that compares the duplicate document against the original document. It relies only on the use of words not on the language specific. Intrinsic Plagiarism is almost similar to the authorship attribution. Intrinsic Plagiarism detection uses the 'n-gram profiles' to compare with the whole document. First, it removes the numbers and also the unwanted letter except the a-z word. Second, it removes all the stop-words and then converts all words into lowercase. Next, using a word frequency based algorithm a vector v is built for all the words in the document. Then the complete document will collected as a cluster c. For each segment or group a frequency vector is computed. Let V be a vector of words that defines word w, as a basic unit of discrete data, indexed by {1. . . |V|}. A document d is a sequence of S words (|d| = S) defined by w= (w1, . . . ,wS), where ws is the sth word in the message. Finally, a corpus is defined by a collection of D documents denoted by C = (w1, . . . ,w|D|).These method segments documents according to stylistic inconsistencies and decide whether or not a document is plagiarism-free. A set of heuristic rules is introduced that attempt to detect plagiarism on either the document level or the text passage level as well as to reduce the effect of irrelevant stylistic changes within a document.

External Plagiarism Detection:

External Plagiarism detection is used to compare all the words with the original document. The comparison between the document is made quickly, effective whether the document is plagiarized or not. The comparing documents and their outputs are not given in detailed information from which the document is copied. Next, the same document is compared again then the detail information will be given and also it tells from which paragraph it's copied. Sometimes it uses the n-gram profiles or string-matching algorithm to give some flexibility to the detection.

External plagiarism detection is used to execute the search space reduction method and also to find plagiarism passage. The search space method aims at quickly identify those pair of documents that potentially have some text in common, possibly one of them having plagiarized from the other.

Authorship Attribution

Authorship attribution is similar to the intrinsic plagiarism detection but only few differences in the styles. It's the task of characterizing the writing style of a document. Whenever problem arises regarding documents and there must be clarification. Linguistic feature as been selected for the writing style of a document. Sometimes it includes the syntactical and lexical analysis for the character, word and sentence. Authorship attribution is used in the large data set and also in small data set in some of the verification task.

Working

The preprocessing, where 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. The method uses word unigrams and considers all words; stop-words are not removed. Next, a word-frequency-based algorithm to test the self-similarity of a document is proposed. Then, the complete document is clustered creating groups C. Secondly the term frequency, in this every segment is compared against the whole document only in terms of the words present in the segment. Finally, all segments are classified according to their distance with respect to the document's style.

Third module describe about the stylometric frequent pattern that can be extracted and mined along with lexical character features, lexical word features and syntactical features. The proposed work detects the plagiarism for documents based on collecting author profiles of documents. Different documents of same author in the various domains can be taken. To be able to distinguish different authors within the same document, writing style present in the text are characterized. Then the features from these documents such as stylometric frequent pattern are extracted. Stylometric is a form of authorship recognition that relies on the linguistic information found in a document. The extracted frequent pattern of features identifies the author of the source document in which the document under test is being plagiarized.

Fourth module describe about the extracted features can be classified by using Machine learning algorithm called Support Vector Machines method (SVM). The Support Vector Machines method (SVM) achieves predominantly the best results in comparisons of machine learning approaches to the authorship recognition problem. Finally the proposed method uses combined approach of text or word of synonyms and stylometric feature can be done to estimate the plagiarism of document in terms of frequency of words. The above mentioned process efficiently detects plagiarism rather than existing work.

Fifth module describe about the naïve bayes classifier and the decision tree learning algorithm. Using these two algorithms it tells about the time and accuracy of the extracted features. Decision tree is the best algorithm where it reduces the time and accuracy.

Conclusion

Intrinsic plagiarism detection and External plagiarism detection can be used for plagiarism detection by comparing the documents. For doing plagiarism detection not all the possible source is available. The idea to analyze the document looking for variations in the writing style. The proposed method uses combined approach of text or word of synonyms and stylometric feature can be done to estimate the plagiarism of document in terms of frequency of words. Decision tree is to reduce the time and to improve the accuracy.

References

[1] *Kasprzak, J., & Brandejs, M. (2010). Improving the reliability of the plagiarism detection system – lab report for pan at clef 2010. In M. Braschler, D. Harman, & E. Pianta, (Eds.), CLEF 2010 labs and workshops, notebook papers. 22–23 September 2010, Padua, Italy.*

[2] *Stamatatos, E. (2009). Intrinsic plagiarism detection using character n-gram profiles. In B. Stein, P. Rosso, E. Stamatatos, M. Koppel, & E. Agirre (Eds.), SEPLN 2009 workshop on uncovering plagiarism, authorship, and social software misuse (PAN 09) (pp. 38–46). CEUR-WS.org.*

[3] *Seaward, L., & Matwin, S. (2009). Intrinsic plagiarism detection using complexity* *analysis. In B. Stein, P. Rosso, E. Stamatatos, M. Koppel, & E. Agirre (Eds.), SEPLN 2009 workshop on uncovering plagiarism, authorship, and social software misuse (PAN 09) (pp. 56–61). CEUR-WS.org.*

[4] *Stein, B., Lipka, N., & Prettenhofer, P. (2011). Intrinsic plagiarism analysis. Language Resources and Evaluation, 45, 63–82*

[5] *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.*