INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

A SURVEY ON PLAGIARISM DETECTION IN TEXT MINING

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

[1]PG Full Time Student, CSE Department, Karunya University, Coimbatore

[2]Assistant Professor, CSE Department, Karunya University, Coimbatore

[1]rubijesus1@gmail.com, [2]leela@karunya.edu

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.

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

1. 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.

2. Related Works:

2.1 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. As Stein et al. (2011) [4] introduced, multiple writing style characteristics were tested in order to determine plagiarism, for example lexical character features, lexical word features and syntactical features.

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.

Stamatos (2009) [2] presented a method for intrinsic plagiarism detection. The style variation within a document using character n-gram profiles and a style-change. Style Profiles are developed using a sliding window. These n-gram profiles are used for getting information about the writer's style.

Seaward and Matwin (2009) [3] introduced Kolmogorov complexity algorithm is a way of extracting structural information from texts for intrinsic plagiarism detection.

2.2 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.

Kasprzak and Brandejs (2010) [1] introduced their model for automatic external plagiarism detection. It consists of two main phases, building the document index and computing the similarities. It uses word n-grams, with n ranging from 4 to 6, and takes into account the number of matches of those n-grams between the suspicious documents and the source documents for computing the detections.

2.3 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. Baayen, van Halteren, and Tweedie (1996) [5] studies the use of words and the use of syntax-based measures

Authorship Attribution analyzes each documents and it generate the multiple set of features for each document. Next, multiple classification model is made based on the set of features.

3. 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. We explore the problem of text plagiarism and the possibility of its detection by the use of algorithm. We used intrinsic plagiarism algorithm for discovering plagiarism by analyzing only the suspicious document. The character n-gram is used for getting information about the writer's styles.

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.