Text extraction technique applied to plagiarism detection: The semantic analysis of statements for analyzing the writing style

Joshi O D

Assistant Professor, Sharad Institute of Technology, College of Engineering
Yadrav-Ichalkaranji. Maharashtra(India)-416121
Omkar.joshi86@sitcoe.org.in

Pudale A H

Assistant Professor, Sharad Institute of Technology, College of Engineering
Yadrav-Ichalkaranji. Maharashtra(India)-416121
Amol.pudale01@sitcoe.org.in

Ghorpade R D

Assistant Professor, Sharad Institute of Technology, College of Engineering
Yadrav-Ichalkaranji. Maharashtra(India)-416121
jeetghorpade@sitcoe.org.in

Abstract— Detecting the plagiarized data is of most concern to research organizations, industries and educational institutions on the computational systems for such a task has become important. Existing methods for detection of plagiarized document compute the similarity of document-to-document basis. We proposed system that will do text mining, exploring the use of keywords and semantic analysis of statements as a feature for analyzing a document. The main goal is to analyze the statements and keywords, looking for segments of the document that is written by other authors. This is considered as a semantic analysis using keyword based statement analysis where paragraphs with unique in style. This approach does work on the use of keywords and semantic analysis of statements, so it is do not require any language specifications. We feel that this feature shows improvement in this area. It will achieve tremendous results compared to existing models.

Keywords-Plagiarized, semantic analysis, text extraction

I. INTRODUCTION

Plagiarized data is of most concern to research organizations, industries and educational institutions. There are many examples available for copy-paste and plagiarism can turns into big problem. By tremendous growth of contents on the internet, it is easy to find everything. Detection of such cases can become a tedious task. There are various methods need to be implemented, ranging from document analysis to scan the keywords and do the semantic analysis of the statements on the Web, to approaches that utilize language-specific features. Plagiarism detection aims to discovering plagiarized data by analyzing the documents. This is achieved by current algorithms, usually use writing style modeling techniques, looking for useful variations. In the external plagiarism refers to the task of comparing the doubtful documents against possible sources. This is the approach in which the systems starts with some kind of labeling of the source documents, and afterwards look at features and coincidences in document before generating the detected copied passages as a result. From as it is document transfer to paraphrasing, many levels of plagiarism techniques can be used. We proposed a system for writing style, targeted at looking significant deviations in a document's writing style; these differing parts may have been copied or plagiarized, and are probably useful as a starting point to search for possible source candidates. A way of searching over the Internet for sources by taking information from the document; anyone can use these changed parts or segments as inputs for the query building. This paper is structured as follows. First, in Section 2, the definition of the problem is presented. In Section 3 we review plagiarism detection research and approaches. In Section 4, the proposed intrinsic plagiarism detection method is described. In Section 5 the conclusions are discussed.

II. PROBLEM DEFINATION

One can consider various types of plagiarism in available documents. The most commonly criticized and easiest to detect is the infamous copy-paste, or copying literals, found especially in under graduate students' work in academic institutions. There are different types of plagiarism are:

A. Exact Copy. The paragraph is copied same to same, without citation.
B. Paraphrasing. The text is modified, but the concept and part of statements remain same.

C. Plagiarism of ideas. The concept is copied using similar words and various resources like language.

It is very easy to find the exact copying if the source documents are available. Due to the availability of the resources plagiarism becomes more sophisticated due to which difficulty of detecting cases of plagiarism increases, due to modifications, the copied text become more complex to identify. This is due to automatic methods and algorithms for available for plagiarism detection; it utilizes resource like language are simple to use in stream of computer science. But when the concept is copied using a different statements, and even when paraphrasing, the identification becomes harder. It is absolutely hard to normalize the text to capture the words that gives the idea behind it using various algorithms. The problem of copying contents can be easily identified in many areas, thus it affects in multiple ways.

III. RELATED WORK

The plagiarism issue can be treated from two different ways, prevention and detection of plagiarized data. As Gabriel Oberreuter, Juan D. Velásquez states using the words, both combined working and find new approach to detect the plagiarized data from the Web using the words. While copy detection only help after the plagiarism has been detected, prevention techniques can create awareness only to educate and motivate people not to do it. Copy plagiarism detection methods, are easier to invoke, and possible to tackle the problem at different levels, from simple comparisons to complex automatic algorithms. A overview about plagiarism detection approaches is presented.

A. Authorship Attribution

It is the task of writing style of a document, aiming at recognizing the style of a particular author. As Juola (2006) gives extensive review explains, authorship attribution has been important in some of cases, confusion related to documents and their authors must be clarified. In automatic authorship attribution it's important to define and select attributes that can considered in the writing style of different authors. With this concern, nearly all research conducted in automatic authors attribution face and treat this as very basic problem. Grieve (2007) studies and gives different measures; Baayen, van Halteren, and Tweedie (1996) research about the use of words and the use of syntax-based measures. In van Halteren (2004) a ''linguistic profile'' is build depend upon multiple linguistic measures.

B. External plagiarism Detection

This method refers to the task of comparing a suspicious document against the possible sources. If a case is found, it must have relevance with the passage in the suspicious document and in source document. While case of a comparison of various documents, few factors become more important. First, the comparison between documents should be made quickly, basically it consider a huge collection of possible resources. Second, the comparison should be effective, detecting slightly changed passages as well as not changed ones. One can separate the multiple approaches proposed so far into two categories. In this category, it may possible that approaches from the machine learning community are need to consider. The first approach, as it uses the technique to infer the latent semantic relations and subsequent determination of the document similarity. The second category of approaches are the ones that go further and provide complete information, especifically indicating the copied passages found. There are many approaches to find the plagiarism. Author Gabriel Oberreuter, Juan D. Velásquez find the approach of Text mining which was applied to detect plagiarism with the use of words for finding different deviations in writing style.

C. Intrinsic plagiarism Detection

This type of detection refers to the analysis of a document, in which it tries to find if a portion of the text has been plagiarized. The recent concept was introduced by Meyer zu Eißen and Stein (2006), and is close related to authorship attribution, where it analyzes the writing style of the text. In this approach, detection is useful when no references are available or not all the possible source documents. In last few years, more studies have been published in which the intrinsic plagiarism detection problem is more investigated. Stamatatos (2009) gives a method for intrinsic plagiarism detection. This approach attempts to style variation within a document using character profiles and a style-change function depends on dissimilarity measure recently proposed and used for identification of author. Stein et al. (2011) discussed their work a description of features used for modeling of writing style, targeted at plagiarism detection. Their result also included such features to discover plagiarism cases, in which the top three performers were the average number of per word and the utilization of the term ''of''. Oberreuter et al. (2011) reported considerable results by analysis in the use of words.

IV. PROPOSED METHOD FOR INTRINSIC PLAGIARISM DETECTION

In our proposed system, First we will be implementing the text classification algorithm, which will extracting and do the separation of the statement into different words (noun, verbs, etc). After getting the words separately, we apply semantic analysis approach to find out meaning of the statements inputted by user. Also by applying text extraction method, will also extract the keywords from the statements.

While starting of system it starts analysis of the keywords to find the relevance of inputted statement to the other available documents on Web. It's possible that keywords are available frequently in many other documents, so system gives result like the inputted page is plagiarized data. To overcome this problem, we also proposed module that will not only extract the key words but also find meaning of the statement. Because of this, though system found the page as plagiarized but the meaning of the inputted statement and meaning of source statement differ. So the inputted page by user should not be considered as plagiarized page. In our proposed module every segment is compared with whole document only in terms of the keywords present in the segment. If the keywords are present then system will automatically find the meaning of the statement. If meaning of the statement is matched then system generate the result as Plagiarized. This system will also consider that if certain words are typically used in a certain segment, the comparison of that segment with document would low value, because the frequency of those words would be the same in both the complete document and in the passage. Finally, all segments are classified with respect to their distance with respect to the document's style.

The function will going to construct in such way that, segment of the document have many words are exclusively in that area will have a less value. The idea is that the use of words should be stable, with the high ratio of words throughout the document. If a portion of the text has a high availability of its words isolated in that portion, the function value will be below the average, hinting at a possible change in writing skill. The algorithm considers the data of each document to construct and evaluate variations in style; the function remains stable over changing document lengths. The strong assumption here is that the majority of the statements was written with the same writing style, otherwise no reliable information could be extracted from this system.

V. CONCLUSION

In this proposed system, we will going to explore the problem of text plagiarism and the possibility of its detection by the use of various computer terms. By the huge demand and usage of digital documents, the usage of plagiarism is increased. To overcome this problem, Text extraction techniques, semantic analysis techniques are used. System may face the problem of collection of possible resources to compare the suspected documents with and finding the meaning of inputted and available documents. This represent an entire problem and it is very usual that ideal and real sources are not available, limiting the potential of algorithms that compute similarity document to document. Algorithms that not required to rely on the available sources are being studied. This is why so called intrinsic plagiarism detection concept was came into the picture. The idea to analyze the document looking for variations that could hint at plagiarized passages was carefully utilizing different writing style makers are being introduced. The study of linguistic features for the data mining process here will be crucial, therefore the exploration of different approaches and writing style characteristics will always appreciated. The major experiments will have to conduct using documents used in English, but the method does not uses language dependent features such as verbs or stop words, thus providing a starting point to experiment with other languages

ACKNOWLEDGEMENT

The authors would like to acknowledge the continuous support of the Mr Anil Bagane, Exe. Director, Sharad Institute of Technology, College of Engineering.Yadrav-Ichalkaranji. Authors are also would like to thanks to Principal of our institute Dr Sanjay A Khot for continuous support and appreciation.

REFERENCES

[1] Gabriel Oberreuter, Juan D. Velásquez (2013). Text mining applied to plagiarism detection: The use of words for detecting deviations in the writing style in Elsevier.

[2] 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.

[3] 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.

[4] 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.

[5] Berry, M. W., Dumais, S. T., & O'Brien, G. W. (1995). Using linear algebra for intelligent information retrieval. SIAM Review,37, 573–595.

[6] 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.

[7] 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.Ceska, Z. (2008). Plagiarism detection based on singular value decomposition. In GoTAL '08: Proceedings of the sixth

[8] international conference on advances in natural language processing (pp. 108–119). Berlin/Heidelberg: Springer.

[9] 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.

[10] [10] Grieve, J. (2007). Quantitative authorship attribution: An evaluation of techniques. Literary and Linguistic Computing, 22, 251–270.

[11] 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.D. Clough, (Eds.), CLEF 2011 labs and workshop, notebook papers. 19–22September 2011, Amsterdam, The Netherlands.
[12] Grozea, C., & Popescu, M. (2011). The encoplot similarity measure for automatic detection of plagiarism – notebook for pan at CLEF 2011. In V. Petras, P. Forner, & P. D. Clough, (Eds.), CLEF 2011 labs and workshop, notebook papers. 19–22 September 2011, Amsterdam, The Netherlands.
[13] van Halteren, H. (2004). Linguistic profiling for author recognition and verification. In Proceedings of the 42nd annual meeting on association for computational linguistics ACL '04. Stroudsburg, PA, USA: Association for Computational Linguistics.
[14] Jun-Peng, B., Jun-Yi, S., Xiao-Dong, L., Hai-Yan, L., & Xiao-Di, Z. (2003). Document copy detection based on kernel method. In Proceedings of the 2003 International Conference on Natural Language Processing and Knowledge Engineering (pp. 250–255).
[15] Juola, P. (2006). Authorship attribution. Foundations and Trends in Information Retrieval, 1, 233–334.
[16] 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.