Computer Engineering and Applications Vol. 1, No. 1, June 2012

Survey of Text Plagiarism Detection

Ahmed Hamza Osman[1,2], Naomie Salim[1], and Albaraa Abuobieda[1,2]

[1]*Universiti Teknologi Malaysia, Faculty of Computer Science and Information Systems, Skudai, Johor,Malaysia,*
[2,3] *International University of Africa, Faculty of Computer Studies, Khartoum, Sudan*
Tel +60147747409/+607 5532208 Fax +607 5532210
[1]*ahmedagraa@hotmail.com*
[2]*Naomie@utm.my*
[3]*albarraa@hotmail.com*

ABSTRAKSI

Dalam tulisan ini berisi review dan membuat daftar keuntungan dan keterbatasan dari teknik yang efektif signifikan digunakan atau dikembangkan dalam deteksi teks plagiarisme. Ditemukan bahwa banyak metode yang diusulkan untuk mendeteksi plagiarisme memiliki kelemahan dan kekurangan untuk mendeteksi beberapa jenis teks dijiplak. Makalah ini membahas beberapa isu penting dalam deteksi plagiarisme seperti; Tugas deteksi plagiarisme, proses plagiat deteksi dan beberapa teknik deteksi plagiarisme saat ini.

Kata Kunci: Deteksi Plagiarisme, Proses – Proses Deteksi, Teknik Deteksi.

ABSTRACT

In this paper contains review and list the advantages and limitations of the significant effective techniques employed or developed in text plagiarism detection. It was found that many of the proposed methods for plagiarism detection have a weakness and lacking for detecting some types of plagiarized text. This paper discussed several important issues in plagiarism detection such as; plagiarism detection Tasks, plagiarism detection process and some of the current plagiarism detection techniques.

Keywords: Plagiarism Detection; Detection Process, Detection Techniques.

1. INTRODUCTION

There are many types of plagiarism, such as copy and paste, redrafting or paraphrasing of the text, plagiarism of idea, and plagiarism through translation from one language to another. These types have made plagiarism one of the serious problems in academic area precisely. A modern research found that 70% of students confess to a few plagiarism, with about half being guilty of an earnest cheating offence on a written assignment. additionally, 40% of students confess to using the "cut- paste" method when completing their assignments [1]. Differentiating between the plagiarized documents and non-plagiarized documents in an effective and efficient way is one main issue in plagiarism detection field.

According to Carroll [2], at least 10% of student's work is likely to be plagiarized in USA, Australia and UK universities [3]. Current methods of plagiarism detection are based on the characters matching, n-gram, chunks or terms.

The objective of this study is to survey some of the important issues in text plagiarism such as detection tasks, detection process and detection techniques that were proposed to handle plagiarism problem for text documents.

The rest of the paper is organized as follows: Section 2 provides a description of the plagiarism detection tasks. Section 3 discusses the plagiarism detection process. Full descriptions of the underlying idea involved in the some of the current methods are presented in Section 4, whereas Section 7 concludes the paper.

2. PLAGIARISM DETECTION TASKS

The first step in dealing with plagiarism is to clearly define the tasks at hand. Based on Potthast et al., [4], plagiarism detection was divided into two main tasks is:

2.1. Extrinsic Plagiarism Detection Task

Extrinsic plagiarism detection assesses plagiarism on a reference to one or more source documents in the corpus. This task tries to utilize the capability of the computer to search for similar documents inside a corpus and retrieve possibly plagiarized documents. Examples of the studies that have been made concerning this type of task include [5-14].

2.2. Intrinsic Plagiarism Detection Task

Intrinsic plagiarism detection evaluates cases of plagiarism by searching into possible suspicious documents in isolation. This type tries to represent the ability of the human to detect plagiarism by examining differing writing styles. "Intrinsic plagiarism aims at identifying potential plagiarism by analyzing a document with respect to undeclared changes in writing style. Several studies had been made under this task such as [15-19].

3. PLAGIARISM DETECTION PROCESS

Unfortunately, many academic institutes do not take plagiarism as seriously as they should. Often they take an "ostrich" approach and turn a blind eye to any wrong doing or at best they may have a very soft policy against plagiarism equating it with bad behavior. However, more and more institutes are taking plagiarism seriously [20].

2.1. Plagiarism detection process stages

As illustrated below in Figure 1, Lancaster and Culwin [21] define the important stages used for plagiarism detection as collection, analysis, confirmation and investigation. These four stages are important for designing error free process.

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

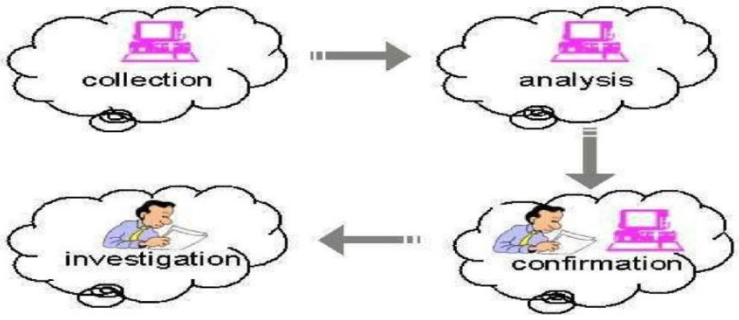

Figure 1. Four-stage Plagiarism Detection Process

In this section, these four stages and their functions will be discussed.

2.1.1. Stage One: Collection

This is the first stage of Plagiarism Detection Process, and it entails the student or researcher to upload their assignments or works to the web engine, the web engine acts as an interface between the students and the system.

2.1.2. Stage Two: Analysis

In this stage all the submitted corpus or assignments are run through a similarity engine to determine which documents are similar to other documents. There are two types of similarity engines, first intra-corpal engine and second extra-corpal engine. The intra-corpal engines work by returning ordered list between each similar pairs. By contrast, the extra-corpal engines return suitable web links.

2.1.3. Stage Three: Confirmation

The function of this stage is to determine if the relevant text has been plagiarized from other texts or to determine if there is a high degree of similarity between a source document and any other document.

2.1.4. Stage Four: Investigation

This is the final stage of a Plagiarism Detection Process and it relies on human intervention. In this step a human expert is responsible for determine if the system ran correctly as well as determining if a result has been truly plagiarized or simply cited.

All four of these stages rely on recognizing the similarity between documents and as a result, they rely on efficient algorithms to search out the similarities between the documents. There is also an element of time complexity required for the human to confirmation and investigation suspected instances of plagiarism.

4. CURRENT PLAGIARISM DETECTION TECHNIQUES

This section mainly discusses some of recently proposed plagiarism detection techniques. These techniques can be classified into character-based methods, structural-based methods, Classification and Cluster-Based Methods, Syntax-Based Methods, Cross language-Based Methods and Semantic-Based Methods and Citation-Based Methods [22].

3.1. Character-Based Methods

3.1.1. Fingerprint

According to Alzahrani's, Survey on Plagiarism Detection Methods [23], common plagiarism detection techniques rely on character-based methods to compare the suspected document with original document. Identical string can be detected either exactly or partially using character matching approaches.

Different plagiarism algorithms adopted the text as character n-gram such as C.Grozea [14], J. Kasprzak [24] and Basile [25] where they used character 16-gram, 8-gram, and 5-gram matching, respectively. In these methods, the degree of similarity between two strings grams depends on the number of identical characters between strings.

Heintze [26], Broder [27] and Monostori et al., [28] proposed a fingerprint method to find the string matching and plagiarism detection based on common fingerprints proportion. These methods obtained good results but failed when the plagiarized part was modified by rewording or changing some words in the suspected text.

3.1.2. String similarity

Brin et al., [29] introduced a plagiarism detection system from Stanford Digital Library Project named Copy Protection System (COPS), which detected document overlap by relying on string matching and sentences. Its main drawback was that it failed to consider individual words and took the whole sentence as one part. The shortcomings of COPS were solved by Shivakumar and Garcia-Molina [30], who developed a new method called Stanford Copy Analysis Method (SCAM) to improve the COPS using Relative Frequency Model (RFM) to emphasize subset copies. FM was an essential asymmetric similarity measure for plagiarism detection. The main advantage of SCAM was that it can find the overlapping similarity between the parts of sentences, but many terms can be misleading in documents sharing comparison. Si et al., [31] proposed a new mechanism for plagiarism detection called CHECK, which was similar to SCAM. Both of these mechanisms adopted information retrieval techniques and worked on overlapping detection based on frequency of words. The CHECK technique, built on an indexed structure known as structural characteristic (SC), used parsed documents for building the SC. It captured plagiarism by focusing on the key word proportion of structural characteristic for the nodes. CHECK covered structured documents only, and ignored the unstructured documents.

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

3.2. Structural-Based Methods

It is worth noting that all the studies above described character-based methods. In fact, all these methods focused on lexical features of the text in the document. Many studies that proposed different methods in field of plagiarism detection focused on structure features of the text in the document such as headers, sections, paragraphs, and references. Tree-Structured Features Representation is one of the recent developments that focused on structure features.

Tree-Structured Features Representation is a rich multi-layer self-organizing maps model (ML-SOM) for text documents [32-33]. Chow and Rahman [10] adopted a tree-structured representation with ML-SOM for information retrieval and plagiarism detection. They built their idea based on two layers, a top layer and a bottom layer. The top layer presented clustering and retrieving of documents while the bottom layer utilized a Cosine similarity coefficient to capture similar and plagiarized text.

3.3. Classification and Cluster-Based Methods

Document clustering technique is one of the information retrieval techniques that were used in many fields such as text summarization [34], text classification [35] and plagiarism detection [36]. It is used to improve the retrieval data using reduction of the searching time in document location for text summarization and reduce the comparison time in plagiarism detection. Another approach by Si et al., [31] and Zini et al., [37] uses specific words (or keywords) to find similar clusters between documents.

3.4. Syntax-Based Methods

Elhadi and Al-Tobi [7] introduced a duplicate detection technique for syntactical structures of document. This technique looked at using syntactical part-of-speech (POS) tags to represent a text structure as a basis for further comparison and analysis. This technique ordered and ranked the documents using POS tags. Elhadi and Al-Tobi [11] improved the methodology of [7] using Longest Common Subsequence (LCS) to calculate the similarity between the documents and ranked them according to the most relevant extracted documents.

3.5. Cross language-Based Methods

A cross-lingual method for detecting suspected documents plagiarized from other language sources was proposed by [14]. In this method, the similarity between a suspected and an original document was evaluated using statistical models to establish the probability that the suspected document was related to the original document regardless of the order in which the terms appear in suspected and original documents. This approach required the construction of the cross-lingual corpus, which can be a difficult task to compile [14].

3.6. Semantic-Based Methods

Many researchers have done excellent work to calculate the semantic similarity between words between documents using WordNet [38]. Knowledge based measures by Gelbukh [39] identified the semantic similarity between two words by

calculating the degree of relatedness between those words using information from a dictionary or thesaurus. It then makes use of the degree of relationship between those words by examining the word's hierarchy within a thesaurus, such as WordNet, which can be seen as a hand-crafted lexical database. Resnik [40] also used WordNet to calculate the semantic similarity. On the other hand, the Leacock's et al., method counted the number of nodes of the shortest path between two concepts to determine semantic similarity [41].

An improved plagiarism detection scheme based on Semantic Role labeling was introduced by Osman et al., [22]. SRL analyzed and compared text based on the semantic allocation for each term inside the sentence. Osman also introduced a weight or value for each argument generated by SRL to study its behaviours. As a result of studying an argument's behaviours, it was found that not all arguments affect the Plagiarism Detection Process.

3.7. Citation-Based Methods

One of the novel methods in plagiarism detection is a citation-based technique that was proposed by [42]. The method used for identifying academic documents that were read and used without referred to that documents.

According to [43] a citation-based plagiarism detection method is belonged to semantic plagiarism detection techniques because it is focused for detection on semantic contained in the citations used in a text academic documents. It intends to identify similar patterns in the citation sequences of academic works for similarity computation [43].

5. CONCLUSION

In this study the problem of plagiarism detection was considered as it is one of the most publicized forms of text reuse around us today. In particular, it has been shown in this study how the plagiarism problem can be handled using different techniques and tools. However, there are still some weaknesses and shortages in these techniques and tools which will affect the success of plagiarism detection significantly.

ACKNOWLEDGEMENTS

This work is supported by IDF in Universiti Teknologi Malaysia. The authors would like to thank International University of Africa (IUA) and Research Management Centre (RMC) Universiti Teknologi Malaysia for the support and incentive extended in making this study a success.

REFERENCES

[1]. D. McCabe, "Research Report of the Center for Academic Integrity," 2005.
[2]. J. J. G. Adeva, *et al.*, "Applying plagiarism detection to engineering education," 2006, pp. 722-731.
[3]. C. Lyon, *et al.*, "Plagiarism is easy, but also easy to detect," *Plagiary: Cross-Disciplinary Studies in Plagiarism, Fabrication, and Falsification,* vol. 1, 2006.

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

[4]. M. Potthast, *et al.*, "Overview of the 1st International Competition on Plagiarism Detection," in *PAN-09 3rd Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse and 1st International Competition on Plagiarism Detection*, 2009, pp. 1-9.
[5]. R. Yerra, & Ng, Y.-K, "A Sentence-Based Copy Detection Approach for Web Documents," *Fuzzy Systems and Knowledge Discovery*, vol. 3613, pp. 557-570, 2005.
[6]. Z. Ceska, *et al.*, "Multilingual Plagiarism Detection," presented at the Proceedings of the 13th international conference on Artificial Intelligence: Methodology, Systems, and Applications, Varna, Bulgaria, 2008.
[7]. M. Elhadi and A. Al-Tobi, "Use of text syntactical structures in detection of document duplicates," in *Digital Information Management, 2008. ICDIM 2008. Third International Conference on*, 2008, pp. 520-525.
[8]. M. S. A. J. A. Muftah, "Document plagiarism detection algorithm using semantic networks," M.Sc, Faculty Comput. Sci. Inf. Syst. Univ.Teechnol. Malaysia Johor Bahru, 2009.
[9]. A. a. P. R. Barrón-Cedeño, "On Automatic Plagiarism Detection Based on n-Grams Comparison," presented at the Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval, Toulouse, France, 2009.
[10]. T. W. S. Chow and M. K. M. Rahman, "Multilayer SOM with tree-structured data for efficient document retrieval and plagiarism detection," *Trans. Neur. Netw.*, vol. 20, pp. 1385-1402, 2009.
[11]. M. Elhadi and A. Al-Tobi, "Duplicate Detection in Documents and WebPages Using Improved Longest Common Subsequence and Documents Syntactical Structures," presented at the Proceedings of the 2009 Fourth International Conference on Computer Sciences and Convergence Information Technology, 2009.
[12]. M. M. M. Zechner, R. Kern, and M. Granitzer, "External and intrinsic plagiarism detection using vector space models," in Proc. SEPLN, Donostia, Spain2009.
[13]. C.-K. Ryu, *et al.*, "A detecting and tracing algorithm for unauthorized internet-news plagiarism using spatio-temporal document evolution model," presented at the Proceedings of the 2009 ACM symposium on Applied Computing, Honolulu, Hawaii, 2009.
[14]. C. G. C. Grozea, and M. Popescu, "ENCOPLOT: Pairwise sequence matching in linear time applied to plagiarism detection," *Donostia, Spain*, pp. 10-18, SEPLN'09 2009.
[15]. B. Stein, *et al.*, "Intrinsic plagiarism analysis," *Language Resources and Evaluation*, vol. 45, pp. 63-82, 2011.
[16]. S. Meyer zu Eissen, *et al.*, "Plagiarism Detection Without Reference Collections Advances in Data Analysis," R. Decker and H. J. Lenz, Eds., ed: Springer Berlin Heidelberg, 2007, pp. 359-366.
[17]. A. Byung-Ryul, *et al.*, "An Application of Detecting Plagiarism using Dynamic Incremental Comparison Method," in *Computational Intelligence and Security, 2006 International Conference on*, 2006, pp. 864-867.

[18]. E. Stamatatos, "Author identification: Using text sampling to handle the class imbalance problem," *Inf. Process. Manage.*, vol. 44, pp. 790-799, 2008.
[19]. B. Stein, *et al.*, "Plagiarism analysis, authorship identification, and near-duplicate detection PAN'07," *SIGIR Forum*, vol. 41, pp. 68-71, 2007.
[20]. T. Lancaster, "Effective and efficient plagiarism detection," South Bank University, 2003.
[21]. F. Culwin and T. Lancaster, "Plagiarism issues for higher education," *Vine*, vol. 31, pp. 36-41, 2001.
[22]. A. H. Osman, *et al.*, "An Improved Plagiarism Detection Scheme Based on Semantic Role Labeling," *Applied Soft Computing*, 2011.
[23]. S. M. Alzahrani, *et al.*, "Understanding Plagiarism Linguistic Patterns, Textual Features, and Detection Methods," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. PP, pp. 1-1, 2011.
[24]. M. B. J. Kasprzak, and M. K "Finding Plagiarism by Evaluating Document Similarities," *Donostia, Spain*, pp. 24-28, SEPLN'09 2009.
[25]. D. B. C. Basile, E. Caglioti, G. Cristadoro, and M. D. Esposti, "A plagiarism detection procedure in three steps: Selection, Matches and "Squares"," *Donostia, Spain*, pp. 19-23, SEPLN'09 2009.
[26]. N. Heintze, "Scalable document fingerprinting," *USENIX Workshop on Electronic Commerce*, pp. 191-200, 1996.
[27]. A. Z. Broder, "On the resemblance and containment of documents," in *Compression and Complexity of Sequences 1997. Proceedings*, 1997, pp. 21-29.
[28]. K. Monostori, *et al.*, "Document overlap detection system for distributed digital libraries," 2000, pp. 226-227.
[29]. S. Brin, *et al.*, "Copy detection mechanisms for digital documents," *SIGMOD Rec.*, vol. 24, pp. 398-409, 1995.
[30]. N. Shivakumar and H. Garcia-Molina, "SCAM: A copy detection mechanism for digital documents," 1995.
[31]. A. Si, *et al.*, "CHECK: a document plagiarism detection system," presented at the Proceedings of the 1997 ACM symposium on Applied computing, San Jose, California, United States, 1997.
[32]. M. K. M. Rahman, *et al.*, "A flexible multi-layer self-organizing map for generic processing of tree-structured data," *Pattern Recogn.*, vol. 40, pp. 1406-1424, 2007.
[33]. M. K. M. Rahman and T. W. S. Chow, "Content-based hierarchical document organization using multi-layer hybrid network and tree-structured features," *Expert Syst. Appl.*, vol. 37, pp. 2874-2881, 2010.
[34]. M. S. Binwahlan, *et al.*, "Fuzzy swarm diversity hybrid model for text summarization," *Inf. Process. Manage.*, vol. 46, pp. 571-588, 2010.
[35]. V. Mitra, *et al.*, "Text classification: A least square support vector machine approach," *Applied Soft Computing*, vol. 7, pp. 908-914, 2007.
[36]. W.-j. L. Du Zou, Zhang Ling "A Cluster-Based Plagiarism Detection Method," *CLEF (Notebook Papers/LABs/Workshops)* 2010
[37]. M. Zini, *et al.*, "Plagiarism Detection through Multilevel Text Comparison," in *Automated Production of Cross Media Content for Multi-Channel Distribution, 2006. AXMEDIS '06. Second International Conference on*, 2006, pp. 181-185.

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

[38]. C. Fellbaum, "WordNet: An electronic database," ed: MIT Press, Cambridge, MA, 1998.
[39]. S. T. a. A. Gelbukh, "Comparing Similarity Measures for Original WSD Lesk Algorithm," *Advances in Computer Science and Application,* vol. 43, pp. 155-166, 2009.
[40]. P. Resnik, "Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language," *Journal of Artificial Intelligence Research,,* vol. 11, pp. 95-130, 1999.
[41]. C. Leacock, *et al.*, "Using corpus statistics and WordNet relations for sense identification," *Comput. Linguist.,* vol. 24, pp. 147-165, 1998.
[42]. B. Gipp and J. Beel, "Citation based plagiarism detection: a new approach to identify plagiarized work language independently," 2010, pp. 273-274.
[43]. B. Gipp and N. Meuschke, "Citation pattern matching algorithms for citation-based plagiarism detection: greedy citation tiling, citation chunking and longest common citation sequence," 2011, pp. 249-258.