Survey of Plagiarism Detection Methods

Asim M. El Tahir Ali, Hussam M. Dahwa Abdulla, Vaclav Snasel
Department of Computer Science
VSB - Technical University of Ostrava
Czech Republic
asim070@yahoo.com, hussamdahwa@hotmail.com, Vaclav.Snasel@vsb.cz

Abstract — **Plagiarism has become one area of interest for researchers due to its importance, and its fast growing rates. In this paper we are going to survey and list the advantages and disadvantages of the latest and the important effective methods used or developed in automatic plagiarism detection, according to their result. Mainly methods used in natural language text detection, index structure, and external plagiarism detection and clustering –based detection.**

I. INTRODUCTION

With the hug of the information on WWW and digital libraries, Plagiarism became one of the most important issues for universities, schools and researcher's fields. It is so easy through the internet and due to using advanced search engine to find documents or journals by students [27]. Some of the researchers are just copying and pasting others works without reference to the owner of the documents. A good survey of ideas about how to define plagiarism can be found in the [23].

II. DEFINE PLAGIARISM

"Plagiarism, the act of taking the writings of another person and passing them off as one's own. The fraudulence is closely related to forgery and piracy—practices generally in violation of copyright laws" Encyclopedia Britannica [28].

Plagiarism can be considered as one of the electronic crimes, like (computer hacking, computer viruses, spamming, phishing, copyrights violation and others crimes). Plagiarism can be defined as the act of taking or attempting to take or to use (whole or parts) of another person's works, without referencing or citation him as the owner of this work. It may include direct copy and paste, modification or changing some words of the original information from the internet books, magazine, newspaper, research, journal, personal information or ideas.

According to the Merriam-Webster Online Dictionary, to "plagiarize" means:

- To steal and pass off (the ideas or words of another) as one's own.
- To use (another's production) without crediting the source.
- To commit literary theft.
- To present as new and original an idea or product derived from an existing source.
- Also according to Turnitin.com and Research Resources this are considered plagiarism:
- Turning in someone else's work as your own.
- Copying words or ideas from someone else without giving credit.
- Failing to put a quotation in quotation marks.
- Giving incorrect information about the source of a quotation.
- Changing words but copying the sentence structure of a source without giving credit.
- Copying so many words or ideas from a source that it makes up the majority of your work, whether you give credit or not (see our section on "fair use" rules).

Plagiarism can be classified into five categories:

- Copy & Paste Plagiarism.
- Word Switch Plagiarism.
- Style Plagiarism.
- Metaphor Plagiarism.
- Idea Plagiarism.

III. WHY PLAGIARISM DETECTION IS IMPORTANT

In some of the academic enterprises like universities, schools and institutions, plagiarism detection and prevention became one of the educational challenges, because most of the students or researchers are cheating when they do the assigned tasks and projects. This is due to the availability of the resources in the internet. It's so easy to them to use one of the search engines to search for any topic and to cheat from it without citing the owner of the document. So its better and must all academic fields they should have to use plagiarism detection soft-wares to stop or to eliminate students cheating, copying and modifying documents when they know that they will be found.

Some types of plagiarism acts can be detected easily by using some of the recent plagiarism detection soft-wares available on the market or over the internet. However for some of the expert plagiarism who is using some of the anti-plagiarism software which is available over the internet, it needs more efforts to detect the plagiarism or cannot be detected at all.

Plagiarism is practiced not only by student but also there are some staff members who like to publish papers in which some parts are directly copied or partially modified to be one of the famous people.

978-0-7695-4414-4/11 $26.00 © 2011 IEEE
DOI 10.1109/AMS.2011.19

There is a big number of plagiarism soft-wares used for plagiarism detection and many of detection tools have been developed by researchers but still they have some limitations as they cannot prove or they show evidence that the documents has been plagiarized from another document or sources it only shows the similarity and give hints to some other documents. This is if the paper has been published globally in some international journal, but some of universities and some of the research centers still do not taking any action against plagiarism detection which help people to cheat more and more.

So still now by using the recent detection software, plagiarism can not 100% be detected?

Copyrights and legal aspects for use of published documents also can be covered by using plagiarism software, so it can show whether this person has legally or illegally copied the documents or not and it also show the whether this person has permission from the owner to use this document or not.

Plagiarism detection is also one of the most important issues to journals, research center and conferences; they are using advanced plagiarism detection tools to ensure that all the documents haven't been plagiarized, and to save the copyrights from violation for the publishers.

IV. PLAGIARISM DETECTION METHODS

There are a big numbers of methods used for plagiarism detection developed by researchers in past years, here we mentioned the latest and the importance and more effectives tools for automatic plagiarism detection, used by some researchers:

Firstly, some of the researchers use Natural language text copy detection technique, this technique appeared in the 1990s, and has produced three detection approaches [2].

A. Grammar-based method

The grammar-based method is one of the important techniques used for plagiarism detection. It focuses on the grammatical structure of documents, and this method uses a string-based matching approach to detect and to measure similarity between the documents. The grammar-based methods is suitable for detecting exact copy without any modification, but it's not suitable for detecting modified copied text by rewriting or switching some words that has the same meaning. This is considered as one of this method limitations [1].

- Huang [3] proposed a similar web pages detection method based on the LCS (Largest Common Subsequence) algorithm by finding the largest common string between two pages to calculate the similarity of the two pages.
- Schleimer S, D.S.Wilkerson, and A. Aiken [4], Winnowing-Based Text Clustering, used overlapping k-gram method to get hashes of the documents, and obtaining fingerprint by reducing the hashes number in documents, they use the fingerprint for each document that have been obtained in statistic and count the rates of similarity between these documents.
- Hashbreaking [5], DCT [6] are also the grammar-based methods; the only difference between them is how to get the fingerprints of the document.

B. Semantics-based method

The semantics-based method, also considered as one of the important method for plagiarism detection, focuses on detecting the similarities between documents by using the vector space model. It also can calculate and count the redundancy of the word in the document, and then they use the fingerprints for each document for matching it with fingerprints in other documents and find out the similarity. The semantic-based method is suitable for non partial plagiarism as mentioned before use the whole document and use vector space to match between the documents, but if the document has been partially plagiarized it cannot achieve good results, and this is considered as one of the limitations of this method, because it's difficult to fix the place of copied text in the original document [1].

C. Grammar semantics hybrid method

Grammar semantic hybrid method is considered as the most important method in plagiarism detecting for the natural languages. This method, so effective in achieving better and improving plagiarism detection result, is suitable for the copied text including modified text by rewriting or switching some words that have the same meaning, which cannot be detected by grammar-based method. It also solves the limitation of semantic-based method. Grammar semantic hybrid method can detect and determine the location of plagiarized parts of the document, which cannot be detected by semantic-based method, and calculating the similarity between documents. [1].

Secondly, some of approaches used task specific index structures likes:

- Malcolm and Lane [12] used the desktop plagiarism detection system Ferret, which is based on common word tri-grams.
- Basile et al. [13] Encoded texts as a word length sequence and used a downstream vector-based n-gram distance measure for candidate selection.
- Kasprzak et al. [14] Incorporate common text shingles in the pre-selection process and Shcherbinin and Butakov [24] employed hash-based fingerprints for candidate retrieval.

Thirdly, External plagiarism detection method:

The external plagiarism detection relies on a reference corpus composed of documents from which passages might have been plagiarized A passage could be made up of paragraphs, a fixed size block of words, a block of sentences and so on. A suspicious document is checked for plagiarism by searching for passages that are duplicates or near duplicates of passages in documents within the reference corpus. An external plagiarism system then reports these findings to a human controller who decides whether the detected passages are plagiarized or not. A naive solution to this problem is to compare

each passage in a suspicious document to every passage of each document in the reference corpus. This is obviously prohibitive. The reference corpus has to be large in order to find as many plagiarized passages as possible. [7].

This fact directly translates to very high runtimes when using the naive approach. External plagiarism detection is similar to textual information retrieval (IR) Baeza-Yates and Ribeiro-Neto [21]. Given a set of query terms an IR system returns a ranked set of documents from a corpus that best matches the query terms. The most common structure for answering such queries is an inverted index. An external plagiarism detection system using an inverted index indexes passage of the reference corpus' documents.

Such a system was presented in Hoad and Zobel [22] for finding duplicate or near duplicate documents.

Another method for finding duplicates and near duplicates is based on hashing or fingerprinting. Such methods produce one or more fingerprints that describe the content of a document or passage. A suspicious document's passages are compared to the reference corpus based on their hashes or fingerprints. Duplicate and near duplicate passages are assumed to have similar fingerprints.

One of the first systems for plagiarism detection using this schema was presented in Brin, Davis, and Garcia-Molina, [20].

External plagiarism detection can also be viewed as nearest neighbor problem in a vector space Rd.

Examples of this of researches used for external plagiarism detection are:

- Automatic Detection of the Direction of Plagiarism, it was used to determining the direction of the plagiarism, the use an extension of the Encoplot method, tested on large scale of artificial plagiarism, they shown that the on largest plagiarism corpus available to date the problem of direction of the plagiarism is solved with fairly high accuracy (about 75%), but they do not tested in natural language. Critian Grozea and Marius Popescu [11].
- Automatic External Plagiarism detection using passage similarities, this approach used in detecting external plagiarism for the pre-processing stage, to indentify non-English documents and translate them onto English, then the index them and retrieve the top documents that are similar to the suspicious. They divide the retrieved documents into passage s which each passage contains twenty sentences, the plagiarism is detected by identifying the number if overlapped words between suspicious and source passage. Clara Vania and Mirna Adriani [15].
- Sobha Lalitha Devi, Pattabhi R K Rao, Vijay Sundar Ram and A Akilandeswari, developed algorism to detect external plagiarism in PAN-10 competition. The algorithm has two steps 1. Identification of similar documents and the plagiarized section for a suspicious document with the source documents using Vector Space Model (VSM) and cosine similarity measure and 2. Identify the plagiarized area in the suspicious document using Chunk ratio. But the preprocessing of the documents is not done [16].
- Markus Muhr, Roman Kern, Mario Zechner, and Michael Granitzer [17], they present their hybrid system for the PAN challenge at CLEF 2010 [16]. Their system performs plagiarism detection for translated and non-translated externally as well as intrinsically plagiarized document passages; the external plagiarism detection approach is formulated as an information retrieval problem, using heuristic post processing to arrive at the final detection results.
- Zechner et al. The team employed a standard model of textual IR for candidate selection. The source documents were indexed on a sentence level and sentences of suspicious documents were used as queries. Similarity was calculated via the well established cosine measure [18].

Lastly and which the important method to us is clustering in plagiarism detection:

Document clustering is one of the important techniques used by information retrieval in many purposes; it has been used in summarization of the documents to improve the retrieval of data by reducing the searching time in locating the document. It is also used for result presentation. Document clustering is used in plagiarism detection to reduce the searching time. But still now in clustering there are some limitations and difficulties with time and space. [19].

One of the approaches used with clustering is the fingerprint-based approach used to analyze and summarize collection of document and create a kind of fingerprint for it. Some of numerical attributes can be used by fingerprint that somehow reflects in the structure of the document. So by creating fingerprint for each document with some of numerical attributes for each document in the collection, we can easily find the matching or the similarity between documents.

Winnowing was presented with the objective of plagiarism detection, but the fingerprint construction guarantees also a set of theoretical properties in terms of fingerprint density and sub-string matching detection.

Some example of it was done by:

- A cluster-Based Plagiarism Detection method uses the grammar-based method, (Winnowing's fingerprint extraction algorithm) which is divided into three steps: first step is called pre-selecting, narrow the scope of detection using the successive same fingerprint. The second step, is called locating, it is to find and merge all fragments between two documents using cluster method. The third step is called post-processing it deals with some merging errors. Du Zou, Wej-jaining long and Zhang Lin [9].
- Evaluation of Text Clustering Algorithms with N-Gram-Based Document Fingerprints, they have

implemented two traditional clustering algorithms with document representation based on winnowing fingerprints, adapted the similarity measures for working with multi-sets and designed a new way of centroid computation. They compared the performance of winnowing fingerprints with term frequency and mutual information and n-fingerprints with four different metrics and with three different collections [8].

V. CONCLUSION AND FURTHER WORK

Plagiarism is so difficult to be 100% detected by recent methods so it will continue rising and raising up, according to the above survey, each method has some advantages and disadvantages. Most of them use clustering as techniques of sorting and summarization tool. According to the latest research it is advised to use the cluster based retrieval or clustering to achieve better results.

With limitation of the grammar-based method and the Semantics-based method, we suggest that we use Semantics-based method for cluster based method as it will achieve much better results.

For further work, we are going to make some comparisons between the recent software used for plagiarism detection and the above mentioned methods according to: 1- Supported languages, 2- Extendibility, 3- Presentation of results, 4- Usability, 5- Exclusion of template code, 6- Exclusion of small files, 7- Historical comparisons, 8- Submission or file-based rating, 9- Local or web-based, 10- Open source.

REFERENCES

[1] Bao Jun-Peng, Shen Jun-Yi,Liu Xiao-Dong,Song Qin-Bao, "A Survey on Natural Language Text Copy Detection",Journal of Software, 2003, vol.14, No.10, pp.1753-1760.

[2] Wang Tao, Fan Xiao-Zhong, Liu Jie, "Plagiarism Detection in Chinese Based on Chunk and Paragraph Weight",in Proceedings of the Seventh International Conference on Machine Learning and Cybernetics, Kunming, pp.2574-2579, July 2008.

[3] Huang Lian'en, "On the Technologies for Building and Accessing a Web Archive", PhD thesis, Peking University, 2008.

[4] Schleimer.S, D.S.Wilkerson, and A. Aiken, "Winnowing: Local Algorithms for Document Fingerprinting", in Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data, ACM New York, NY, USA, pp.76–85, 2003.

[5] Abdel-Hamid, B.Behzadi, Stefan Christoph, Monika Henzinger, "Detecting the Origin of Text Segments Efficiently", WWW 2009, Madrid, Spain, pp.61-70, 2009.

[6] J. Seo and W. B. Croft, "Local Text Reuse Detection", SIGIR'08, pp.571–578. ACM, 2008.

[7] Mario Zechner, Markus Muhr, Roman Kern and Michael Granitzer, External and Intrinsic Plagiarism Detection Using Vector Space Models, In: Proceedings of the SEPLN'09 Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse. 2009, pp. 47-55

[8] Javier Parapar, Alvaro Barreiro: Winnowing-based text clustering. CIKM 2008:1353-1354.

[9] Du Zou, Wei-jiang Long, Zhang Ling: A Cluster-Based Plagiarism Detection Method - Lab Report for PAN at CLEF 2010. CLEF (Notebook Papers/LABs/Workshops) 2010.

[10] Ameera Jadalla, Ashraf Elnagar: PDE4Java: Plagiarism Detection Engine for Java source code: a clustering approach. IJBIDM 3(2):121-135 (2008).

[11] Cristian Grozea, Marius Popescu: Who's the Thief? Automatic Detection of the Direction of Plagiarism. CICLing 2010:700-710.

[12] Malcolm, J.A., Lane, P.C.R.: Tackling the PAN'09 external plagiarism detection corpus with a desktop plagiarism detector. In: Proceedings of the SEPLN'09 Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse. pp. 29–33 (2009).

[13] Basile, C., Benedetto, D., Caglioti, E., Cristadoro, G., Degli Esposti, M.: A plagiarism detection procedure in three steps: Selection, matches and "squares". In: Proceedings of the SEPLN'09 Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse. pp. 19–23 (2009).

[14] Kasprzak, J., Brandejs, M., Kripac, M.: Finding plagiarism by evaluating document similarities. In: Proceedings of the SEPLN'09 Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse. pp. 24–28 (2009).

[15] Clara Vania, Mirna Adriani: Automatic External Plagiarism Detection Using Passage Similarities - Lab Report for PAN at CLEF 2010.

[16] Sobha Lalitha Devi, Pattabhi R. K. Rao, R. Vijay Sundar Ram, A. Akilandeswari: External Plagiarism Detection - Lab Report for PAN at CLEF 2010.

[17] CLEF (NotebookPapers/LABs/Workshops) 2010.

[18] Markus Muhr, Roman Kern, Mario Zechner, Michael Granitzer: External and Intrinsic Plagiarism Detection Using a Cross-Lingual Retrieval and Segmentation System - Lab Report for PAN at CLEF 2010.

[19] CLEF (Notebook Papers/LABs/Workshops) 2010.

[20] Zechner, M., Muhr, M., Kern, R., Michael, G.: External and intrinsic plagiarism detection using vector space models. In: Proceedings of the SEPLN'09 Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse. pp. 47–55 (2009).

[21] K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. ACM Computing Surveys, 31(3):264-323, 1999.

[22] Brin, S. and Davis, J. and Garcia-Molina, H. (1995) Copy Detection Mechanisms for Digital Documents. In: ACM International Conference on Management of Data (SIGMOD 1995), May 22-25, 1995

[23] Baeza-Yates, R. & Ribeiro-Neto, B. Modern Information Retrieval. Addison-Wesley, 1999.

[24] Timothy C. Hoad and Justin Zobel. Video similarity detection for digital rights management. In Proceedings of the 26th Australasian computer science conference - Volume 16 (ACSC '03), Michael J. Oudshoorn (Ed.), Vol. 16. Australian Computer Society, Inc., Darlinghurst, Australia, Australia, 237-245. 2003.

[25] Hermann A. Maurer, Frank Kappe, Bilal Zaka: Plagiarism - A Survey. J. UCS 12(8): 1050-1084 (2006)

[26] Shcherbinin, Vladislav and Sergey Butakov (2009). "Using Microsoft SQL Server platform for plagiarism detection". In: SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09), p. 36.

[27] Plagiarism. http://www.plagiarism.org/ (last access February 7, 2011)

[28] Encyclopedia Britannica, http://www.britannica.com/EBchecked/topic/462640/plagiarism