Plagiarism Detection over the Web: Review

Mauli Joshi
II[nd] year, Mtech
PDM College Of Engineering For Women,
Haryana, India

Kavita Khanna
Associate Professor
PDM College Of Engineering For Women,
Haryana, India

ABSTRACT
Plagiarism has many meaning depending upon the seriousness of the task. It is piracy of content in the academic conduct and is marked as equivalent to a crime leading to disruption of reputation or much worse suspension. There are many examples from acclaimed universities to much publicized personalities those have been accused for plagiarism. This paper discusses various techniques and methods that have been adopted to detect and prevent plagiarism in articles, journals, scientific publications and the future perspective.

General Terms
Information retrieval, Stop words

Keywords
Plagiarism; Plagiarism Detection; Plagiarism prevention; Web mining.

1. INTRODUCTION
A professional, author or student when either uses other person's work without his permission or present it under his own name is considered as an act of Plagiarizing. This type of stealing is also known as copyright confliction or text reuse. It not only exists in scientific journals but also in field of journalism, arts, over web, on blogs and websites. One way to deal with is to cite the content so that the original author gets the contribution with the proper attribution. With the advent of internet it has become much easier to find information and use it as your own, in some fields like news articles or journals it is not even considered unethical to reuse text, which is not a right thing technically and so should be avoided. The paper is arranged into sections further describing plagiarism its types and techniques to handle it.

2. CLASSIFICATION
Plagiarism is the practice of taking someone else's work or ideas and passing them off as one's own without citing the text, means the author of the original work is not contributed . The whole classification is broadly categorized into the intentional and unintentional plagiarism, all the other types fall in as under. Intentional is when author knowingly plagiarize; these are also described in the figure below, Figure 1. , Types of Plagiarism are: *Direct plagiarism*, is when author cut copies the content to use it as his own. *Paraphrasing*, is when the text is reordered or rearranged but still means the same. *Insufficient acknowledgement plagiarism*, when proper citations are not done in the content. *Mosaic plagiarism*, happens when author doesn't bothers or ignores about his work to be plagiarized because of lack of knowledge or ignorance. *Patchwork plagiarism*, when author copies parts of original work to make his own. *Idea plagiarism*, when author steals the idea of someone else without attributing [1, 2, 3].

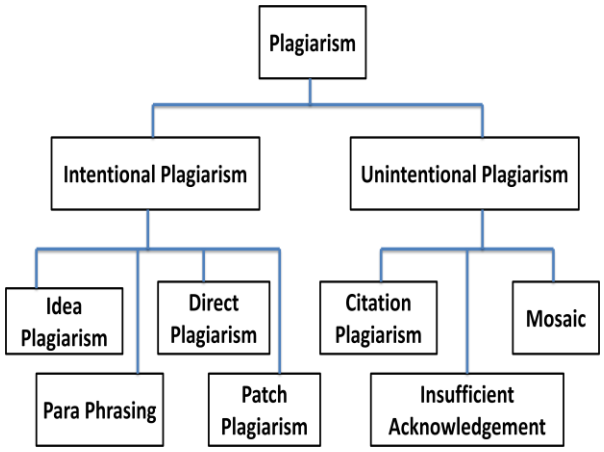

Figure 1: Classification of plagiarism

Example of plagiarism:

Original sentence "University has ratified to students, whether knowingly or not, plagiarism will result as a punishable offence".

Plagiarized sentence" It has been acknowledged by the university that plagiarism is punishable".

It is to be noted that to plagiarize content, replacing words with synonyms is usual, the stop words that are very frequent words, can be used to detect the similarities between original and suspicious document. In the above example, words like 'the', 'that', 'is', 'by' are examples of stop words. Stop words are very useful in finding the plagiarism of attribution [17]. In information retrieval there are similarity measures that are used to capture the similarly between two documents such purity, accuracy, F-measure [10] etc. in order to find relevance precision and recall measures are effective.

Precision = (relevant items retrieved)/ (retrieved items) [18]
Recall = (relevant items retrieved)/ (relevant items) [18]

Both of these vary between values of zero to one, when precision score is 1.0 means that every result retrieved by search is relevant. If recall score is 1.0 means that all relevant documents were retrieved by the search.

2.1 Tools for detecting Plagiarism
Different plagiarism tools have been devised till date, many exist online to help teachers and researchers, some are paid versions and some free to use. Turtnitin [4], Copyscape [5], PlagTracker[6], Viper[7], PlagSpotter [8] are some examples of such tools that takes the original document and makes a check to its existing database or across web to see if its copied. All these are effective in avoiding piracy of

documents and words. The common method that is employed by the online tools is by checking the text on web to detect copied content (takes the measures of information retrieval) and, like Google, it is done by assigning rank and the low page rank denotes duplicated content. These tools uses set of algorithms to find modified text. The source used by these tools is either the internet or the documents submitted to its own database.

2.2 Methods for Detecting Plagiarism

Most of the existing work uses different approaches for plagiarism identification like exact match, sentence based match, finger printing, substring matching. Finger printing is a computer assisted technique, finger print here represents the digest of document which is compared to detect suspicious chunk of data. In Substring matching pair of strings are matched and these substrings are represented on a suffix tree, then the algorithm is applied to detect plagiarism, another method Stylometry identifies the attribution of authorship and is used to capture author's unique writing style. Citation based pattern analysis keeps a check on citation and references used in the text document [19].

3. RELATED WORK

In Year 2011, Salha Alzahrani et al in paper " iPlag: Intelligent Plagiarism Reasoner in Scientific Publications"[9] proposed iPlag which works by combining several analytical procedures. In this framework, approaches like Relevance ranking (R-RANK) and plagiarism screening (P-SCREEN) are adjusted to incorporate citation evidences, structural weights, syntax-based and semantic-based methods into the existing plagiarism detection systems.

In Year 2011 Efstathios Stamatatos et al, in paper," Plagiarism Detection Based on Structural Information" [10] described a method for detecting plagiarized passages in document collections . Author showed that how stopword n-grams are able to capture local syntactic similarities between suspicious and original documents. Also, an algorithm for detecting the exact boundaries of plagiarized and a source passage is proposed.

In Year 2011, Bela Gipp in paper," Comparative Evaluation of Text- and Citation-based Plagiarism Detection Approaches using GuttenPlag" [11] proposed a new approach to identify similar and plagiarized documents based on the citations used in the text and it was shown that citation-based plagiarism detection performs better than the text-based procedures in identifying strong paraphrasing, translation and some idea plagiarism and detection rates can be improved by combining citation-based with text-based plagiarism detection.

In Year 2007, Chi Hong et.al in the paper, "Natural Language processing approach to automatic plagiarism detection" [12] and goal of this approach was to identify copied contents even after intentional changes in the sentence structure and word replacement. Although the syntactic and semantic methods can assist in understanding the meaning of a sentence, instead of its surface structure, the time taken to process sentences is longer than that of comparing sentences at the surface level. Thus, it is suggested that this approach can be applied to detect plagiarism when the source is very likely to be copied. For example, the past assignments in a university may be likely to be copied by the students in the same university.

In Year 2011, Tomas Kucecka performed a work," Obfuscating Plagiarism Detection - Vulnerabilities and Solutions"[13] in which author described the most common ways on how to deceive the plagiarism detection software by introducing four obfuscation categories. Author take several existing plagiarism detection tools and test their resistance against simple but effective obfuscations, proposing method and implementing it into a plagiarism detection system author identifies obfuscated documents.

In Year 2006, Chao Liu et al performed a work," GPLAG: Detection of Software Plagiarism by Program Dependence Graph Analysis" [14] in which author developed a new plagiarism detection tool called GPlag, this tool detects plagiarism by mining program dependence graphs (PDGs). To make GPlag scalable to large programs a statistical lossy filter is proposed and experiment study shows that GPlag is both effective and efficient it detects plagiarism that is not easily detectable by other existing tools.

In Year 2007, Romans Lukashenko et al in paper," Computer-Based Plagiarism Detection Methods and Tools: An Overview" [15] described plagiarism problem and the ways on how to reduce plagiarism, plagiarism prevention and plagiarism detection are discussed. Also widely used plagiarism detection methods are described and the most known plagiarism detection tools are analyzed.

4. PLAGIARISM DETECTION

Plagiarism detection could be done by many techniques one of which is manual detection; however for larger text document it is not efficient. Apart from manual detection, now a days plagiarism detection tools are used in which user can check and compare the work over internet. These tools are more helpful because we now can check it on basis of syntax and semantics and also on the source code. However plagiarism detection in source code is very difficult to capture, graph dependence analysis can be somewhat effective to notice core part of program. Source code plagiarism is plagiarism of programming and it is very easy for someone to use the code as a whole or in modules from original programming to a modified one without being caught. Program Graph Dependence Analysis (PGD) can be used to catch plagiarism in line of codes, which is a graphical representation of source code by vertices and edges. For altering the code plagiarist needs to have ample knowledge which also it increases the amount of effort and the cost of restructuring the code that is not worth of plagiarizing [14]. In academia or in journals tools like turnitin (and many more) can be used to find if any redundant data exists in text and if there is then proper citation is needed to be done.

In figure 2, the process of plagiarism detection has been shown, user enters a document or text at his device to run a check of whether it matches to someone else's work or not. So, source document is retrieved and analysis is performed between original and suspicious document.

Techniques like finger printing, substring matching to analyze the text, in the next phase that is matching the entered text is matched across the web to find corresponding match, if that exists. And the result is shown at the user's site.

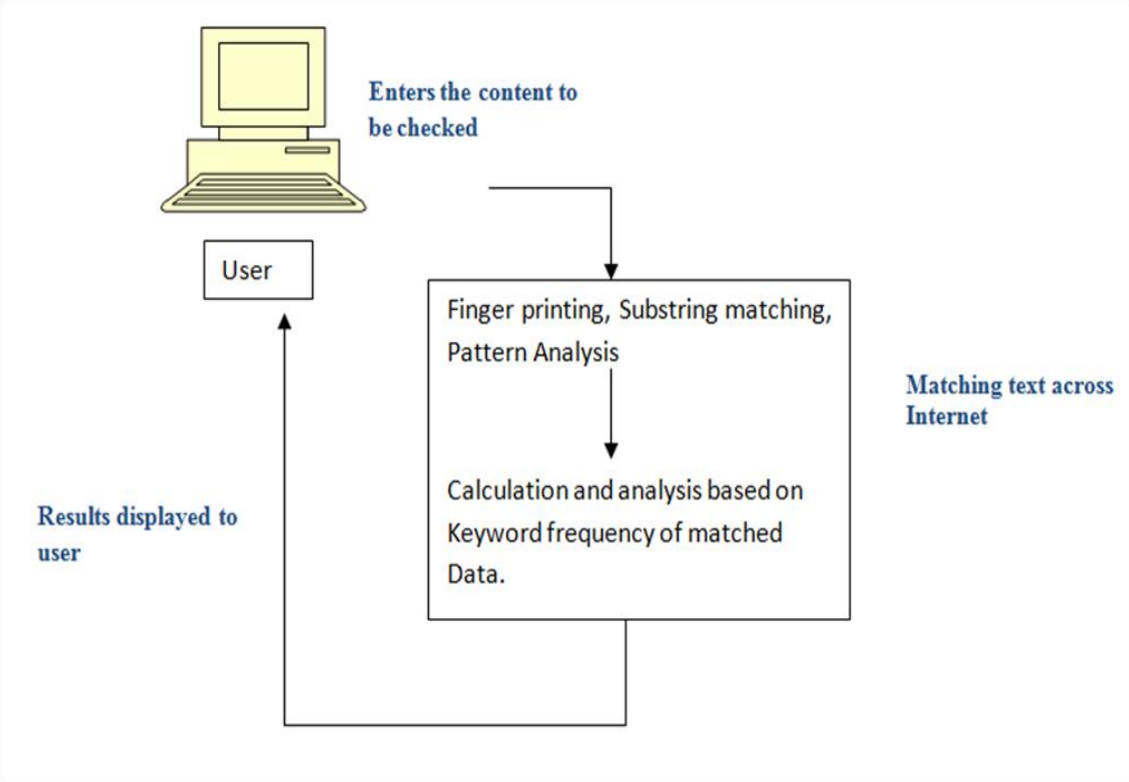

Figure 2: Plagiarism Detection Process

4.1 Web Mining

Web mining is a type of data mining technique in which knowledge is extracted from Web data, Web documents, and hyperlinks between documents.

It is further divided in to web content mining, web usage mining and web structure mining. With relevance to plagiarism detection, involves steps: text extraction, analyzing keyword frequency and presenting with similarity ratio with matched content. For this web content mining is used for information retrieval, extracting association patterns, clustering of web documents and classification of Web Pages.

Similarity measures are used to represent similarities between documents. *Purity* gives fraction of overall cluster size. Each cluster is assigned to the class which is most frequent in the cluster, and then the accuracy of this assignment is computed by counting the number of correctly assigned documents and dividing by *N*. Formally Purity is calculated as below [16]:

$$Purity(\Omega, ℂ) = \frac{1}{N}\sum_{k} max_j \mid \omega_k \cap c_j \mid$$

Where $\Omega = \{\omega 1, \omega 2, \ldots, \omega K\}$ is the set of clusters and $ℂ = \{c1, c2, \ldots, cJ\}$ is the set of classes. ωK is the set of documents in ωK and c_j is the set of documents in c_j. High purity is can be easily achieved when the number of clusters is large; purity is 1 if each document gets its own cluster [16].

Accuracy: is the fraction of clusters that are correct (i.e. it measures the percentage of decisions that are correct) and depicts the fraction of clusters in the dominant category.

5. PLAGIARISM AVOIDANCE

While writing the text best thing to avoid plagiarism could be citation or referencing the text while taking the main idea from source text, ideas and writings used should be acknowledged. And one should use own words while expressing idea. Also there are number of software that could be used to check the content to avoid plagiarism and its consequences.

6. CONCLUSION

This paper discusses about plagiarism, detection techniques and tools and how to avoid plagiarism. Although there are the methods that detect plagiarism but none of them provides full protection, techniques like paraphrasing and modifications are hard to detect by any of them. Also when text is used by converting active to passive voice it is almost impossible to be caught under scanner, but if the source text is divided into smaller strings or chunks it is much easier to detect faulty one as we can separate the text from stop words.

With the information is becoming easily available on electronic media it has become easier to plagiarize but this electronic means has also made to detect the plagiarism at much ease. Numerous tools are available over web to check and analyze the content. By analyzing some of these tools it is known that these are not as good as manual detection method but manual detection also has its limitations.

In future focus should also be shifted to multi lingual plagiarism detection as plagiarist could reuse the source from other language to their own. Like in Hindi it is harder to detect plagiarism because certain words have more than two

different meaning, so it would always be a problem while translating

7. REFERENCES

[1] C Bambaum, Plagiarism, A Student's Guide to recognizing it and avoiding it. Retrieved from http://ww2.valdosta.edu/~cbarnbau/personal/teaching_MISC/plagiarism.

[2] Loveleena Rajeev, Different types of plagiarism, July 18, 2012. Retrieved from http://www.buzzle.com/articles/different-types-of-plagiarism.html

[3] Avoiding Plagiarism, Uefap, Retrieved from http://www.uefap.com/writing/plagiar/plagiar.htm.

[4] Turnitin. Available http://www.turnitin.com

[5] Copyscape. Available http://www.copyscape.com

[6] Plagtracker. Available http://www.plagtracker.com

[7] Viper. Available: http://www.scanmyessay.com/

[8] PlagSpotter. Available http://www.plagspotter.com/

[9] Salha Alzahrani et al," iPlag: Intelligent Plagiarism Reasoner in Scientific Publications", 2011 World Congress on Information and Communication Technologies 978-1-4673-0125-1@ 2011 IEEE (pp 1-6)

[10] Efstathios Stamatatos et al," Plagiarism Detection Based on Structural Information", CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK. ACM 978-1-4503-0717-8/11/10 (pp 1221-1230).

[11] Bela Gipp et al, "Comparative Evaluation of Text- and Citation-based Plagiarism Detection Approaches using GuttenPlag", JCDL'11, June 13–17, 2011, Ottawa, Canada. Copyright 2011 ACM 978-1-4503-0744-4/11/06 (pp 255-258).

[12] Chi-hong Lueng,yuen-Yan, "A Natural Language Processing Approach to Automatic plagiarism Detection",2007, proceeding of 8th ACM SIGITE Conference

[13] Tomas Kucecka et al," Obfuscating Plagiarism Detection - Vulnerabilities and Solutions", International Conference on Computer Systems and Technologies - CompSysTech'11 CompSysTech'11, June 16–17, 2011, Vienna, Austria. ACM 978-1-4503-0917-2/11/06. (pp 423-428)

[14] Chao Liu et al," GPLAG: Detection of Software Plagiarism by Program Dependence Graph Analysis", Industrial and Government Applications Track Paper KDD'06, August 20–23, 2006, Philadelphia, Pennsylvania, USA. Copyright 2006 ACM 1-59593-339-5/06/0008 (pp 872-881).

[15] Romans Lukashenko, Vita Graudina, Janis grundspenkis, "Computer-based plagiarism detection methods and tools", International Conference on Computer Systems and Technologies - CompSysTech , pp. 40-6, 2007

[16] Introduction to Information Retrieval Christopher et al, Evaluation of clustering. Retrieved from http://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-clustering-1.html.

[17] Arun, R., Suresh. V., and Madhavan, C.E.V. 2009. Stopword graphs and authorship attribution in text corpora. In Proceedings of IEEE International conference on semantic computing, 192-196.

[18] Introduction to Information Retrieval Christopher et al, Evaluation of Unranked Retrieval Sets. Retrieved from http://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf pg 155.

[19] Plagiarism Detection, December 2010. Retrieved from http://en.wikipedia.org/wiki/Plagiarism_detection#Detection_methods.