International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163
Volume 1 Issue 7 (August 2014) http://ijirae.com

Plagiarism Detection-Different Methods and Their Analysis: Review

S.A.Hiremath M.S.Otari

Abstract—Plagiarism is possible in every field of day to day life such as plagiarism in music, paintings, pictures, maps, technical articles and drawings, etc. In this paper an overview of different plagiarism detection methods used for text documents have taken. The different plagiarism detection techniques such as text based, citation based and shape based are discussed in brief and compared with respect of their features and performance. Along with this, the paper focuses on different software used for plagiarism detection.

Keywords —Plagiarism, similarity detection, Plagiarism Detection, Citation Analysis, flowchart, image comparison

I. INTRODUCTION

The advancement of information technology and use of smart phones increased the availability of information. Plagiarism 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.[1].

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.

Plagiarism can be classified into five categories: Copy & Paste Plagiarism, Word Switch Plagiarism, Style Plagiarism Metaphor Plagiarism, Idea Plagiarism [2].

There are two types of plagiarism that occurs most frequently:

1. Textual plagiarisms: this type of plagiarism usually done by students or researchers in academic enterprises, where documents are identical or typical to the original documents, reports, essays scientific papers and art design.[4]

2. A source code plagiarism: also done by students in universities, where the students trying or copying the whole or the parts of source code written by someone else as one's own, this types of plagiarism it is difficult to detect [5].

There are many plagiarism detection techniques such as textual based plagiarism, citation based plagiarism, and shape based plagiarism for flowchart. .Textual plagiarism is a type of plagiarism that delivers satisfying results if the plagiarized text is copied (copy & paste), with minor alterations (e.g. shake & paste) or machine translated. However, if the text is paraphrased or translated by a human, the currently used methods yield a very poor performance. Citation-based Plagiarism Detection compares the occurrences of citations in order to identify similarities. The most basic form is to measure the bibliographic coupling strength. strength of the citation based approach lies in identifying translation- and idea-plagiarism or disguised paraphrasing.[3]

Shape based plagiarism for flowchart presents a method for detecting flow chart figure plagiarism based on shape-based image processing and multimedia retrieval. The method managed to retrieve flowcharts with ranked similarity according to different matching sets. However it is unable to identify plagiarism for different figures and charts along with their contents.

II. TEXT BASED PLAGIARISM

This type of plagiarism 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. This 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. It may include copy and paste, modification or changing some words of the original information from the internet book magazine, newspaper, research, journal, personal information or ideas [5].

A. Text based plagiarism detection process stages

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) 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.

3) Stage Three Confirmations: 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.

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.[2]

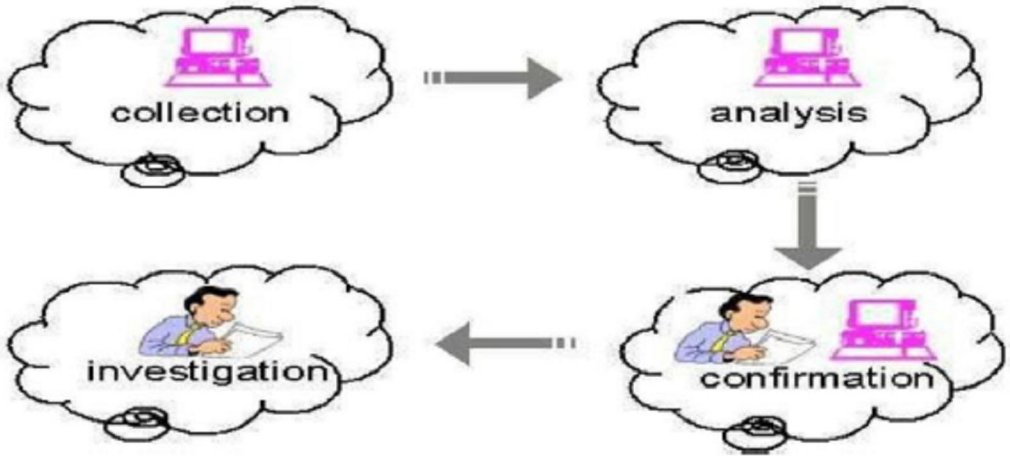

Fig. 1 Four-stage Plagiarism Detection Process[2]

B. Different methods used for textual plagiarism detection

The word for word taking of a phrase, sentence, paragraph(s), or entire source without using quotation marks AND without citing it appropriately. The most prevalent form of plagiarism among high school students. The Internet has made copy & pasting extremely tempting for students 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.

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. 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.

1) *Grammar-based method* : The grammar-based method is important tool to detect plagiarism. 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 is 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

2) *External plagiarism detection method:* The external plagiarism detection relies on a reference corpus composed of documents from which passages might have been plagiarized .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.[5]

III. CITATION-BASED PLAGIARISM

"Citation-based Plagiarism Detection (CbPD) subsumes methods that use citations and references for determining document similarities in order to identify plagiarism".

In the academic environment citations and references of scholarly publications have long been recognized for containing valuable semantic information about the content of a document and its relation to other works. The degree of similarity between citation patterns depends, among others factors, mainly on the amount of shared references (bibliographic coupling strength), and the extent to which the order of included citations, as well as their distance towards each other is similar.

A. Identifying Citation Patterns

Finding similar patterns in the citations used within two scientific texts is a strong indicator for semantic text similarity and the core idea of CbPD. Patterns are subsequence's in the citation tuples CA and CB of two texts A and B that (partially) consist of shared references and are therefore similar to each other.

The degree of similarity between patterns depends on the number of citations included in the pattern, and the extent to which their order and/or the range they cover is alike. Thus, literally matching subsequence's of citations in two documents are a strong indicator for semantic similarity.[3]

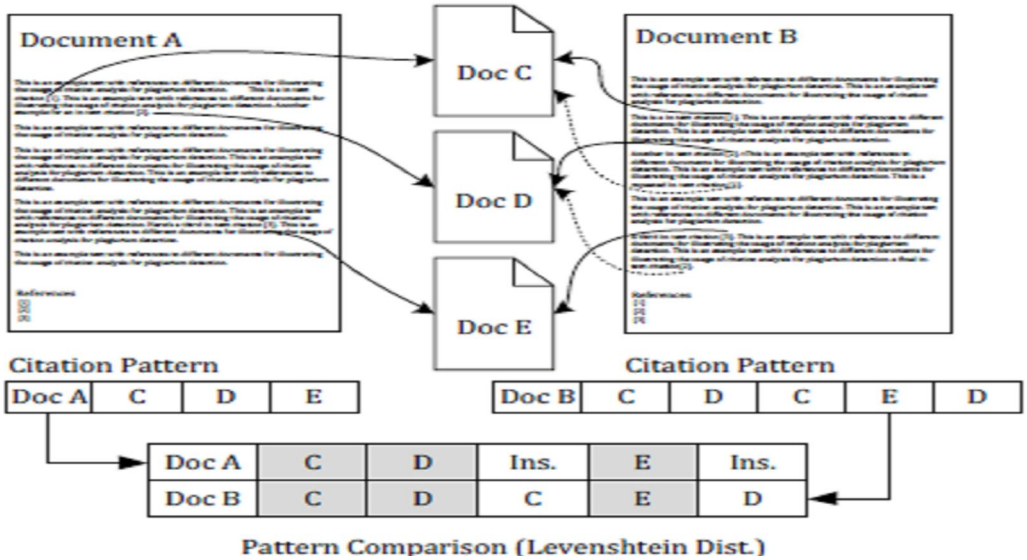

Fig .2 Identifying citation patterns for CbPD[3]

1) *Unknown pattern constituents:* Unlike e.g. in string pattern matching the subsequence's of citations to be extracted from a suspicious text and searched for within an original are initially unknown. Citations that are shared by the two documents are easily identified. However, it is unlikely that all of those shared citations represent plagiarized text passages. For instance, two documents might share 8 citations, of which 3 are contained within a plagiarized text section and 4 are distributed over the length of the text and used along with other non-shared citations without representing any form of plagiarism. The citation sequences of the two documents might therefore look like the following:

Original: 1 2 3 x x x 4 x x 5 x 6 x 7 8

Plagiarism: x x 5 x x x 4 x 3 x 1 x 2 x x 7 x 8

Numbers 1-8 represent shared citations, the letter x non-shared citations. The shared citations 1-3 are supposed to represent a plagiarized passage.[3]

B. CbPDS system architecture

For the Citation-based Plagiarism Detection an Open Source software system in Java coined CitePlag was developed. These steps are performed in our plagiarism detection system:

1. The document is parsed and a series of heuristics applied to process the citations, including their position within the document.
2. Citations are matched with their entries in the bibliography.
3. The citation-based similarity of the documents is calculated.

The developed prototype CbPDS consists of three main components. The first is a Relational Database System (RDBS) termed CbPD database storing data to be acquired from documents as well as detection results. The second is the detection software called CbPD Detector that retrieves data from the CbPD Database, runs the different analysis algorithms to be evaluated and feeds the resulting output back to the CbPD Database. The third component, the CbPD Report Generator, creates summarized reports of detection results for individual document pairs based on adjustable filter criteria. The three-tier-architecture is illustrated in the following figure.

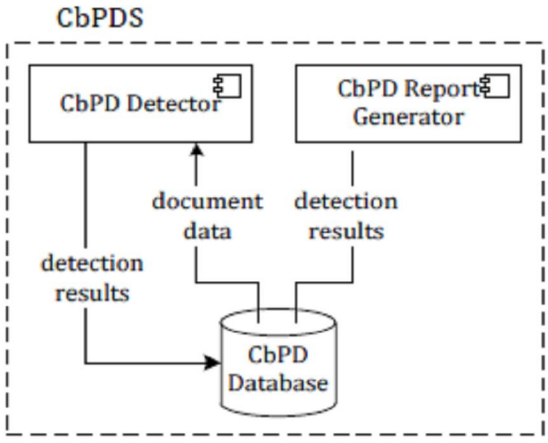

Fig. 3 CbPDS system architecture[3]

IV. SHAPE-BASED PD FOR FLOWCHART

Most, if not all, discard the figures and charts before checking for plagiarism. Discarding the figures and charts results in look holes that people can take advantage. That means people can plagiarized figures and charts easily without the current plagiarism systems detecting it. There are very few papers which talks about flowcharts plagiarism detection.

Therefore, there is a need to develop a system that will detect plagiarism in figures and charts. Flowcharts become a significant issue to explain different kinds of information based on figure types. In some documents, flowcharts are so important to illustrate a lot of details and make it easier to understand methodology of structured design is one of primary steps to build entire system and solving engineering problems that can be explained by using flowcharts and other types of figures.[6]

A . METHODOLOGY

The main goal of this project is to create a figure plagiarism system that is primarily based on shape. This system primarily focuses on flowcharts detection. The database contains flowchart images stored in a single folder. The system will retrieve and rank this database based on a given query by the user. The retrieval system works by detecting shapes in each figure and compare to the shape from the query, as shown in Figure 1.

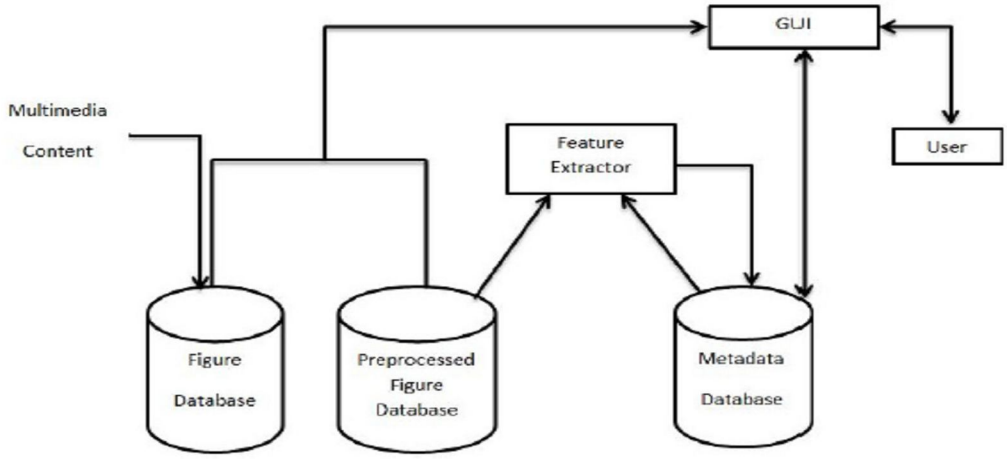

Fig.4 Multimedia Information Retrieval System [6].

1) *Pre-processing:* In order for the system to obtain maximum retrieval result, the obtained figures need to be preprocessed.The pre-processing is done to reduce retrieval errors and help the system accuracy.There are three steps on pre-processing will be done:

1. Thinning
2. Removing connected lines
3. Removing text

2) *Database:* To create the database, first is to group figures and preprocessing figure into separated sets. The databases are created by two sets of databases:
   1. First database is for storing the figures.
   2. Second database is for storing the preprocess figure

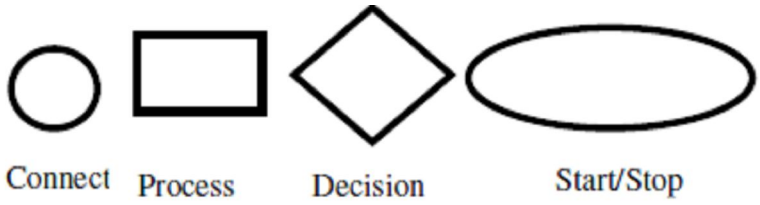

Fig.5 Sample Database Shapes

3) *Metadata Extraction:* From the database the shape can trace back to the metadata. The process of this step is getting the image queries and extracts index shape inside query figures. The same as in database for them metadata each figure has been given an id. The figures can now be stored to the database as well as the metadata, as shown in Table 1.

TABLE I
SAMPLE METADATA DATABASE[6]

FigureID	Connector	StartStop	Decision	Process
1				
2				
3				
4				

V. COMPARISON

The different plagiarism detection techniques such as text based, citation based and shape based are compared based on the two categories: first according to features and secondly according to performance Qualitative comparison used in comparing the features of software, where we are looking for properties of the methods used. Quantitative comparisons used in comparing the performance of the method used.

A. The text-based PDS

It delivers satisfying results if the plagiarized text is copied literally (copy&paste), with minor alterations (e.g. shake&paste) or is machine translated. Text matching approaches continue to be suitable for detecting copy&paste plagiarism, even for short passages. They are also advantageous in that they do not require citation information. The text-based PDS, especially Ferret and WCopyfind, which work with local document comparisons, deliver good results for identifying copy&paste plagiarism given that the sources are available, However, if the text is paraphrased or translated by a human, the text based method yield a very poor performance. Thus they fail to identify e.g. paraphrased, translated and idea plagiarism.

B. Citation-based Plagiarism Detection (CbPD)

It subsumes methods that use citations and references for determining similarities between documents in order to identify plagiarism. compares the occurrences of citations in order to identify similarities. The most basic form is to measure the bibliographic coupling strength (citation overlap). Citation-based Plagiarism Detection is by no means a replacement for the currently used text-based approaches.

CbPD must be carefully verified by humans, especially in cases where only a few citation Whereas the strength of existing PDS lies in detecting plagiarism on the sentence level in the form of identifying similar or identical consecutive words, the strength of the citation based approach lies in identifying translation- and idea-plagiarism or disguised paraphrasing Citations and citation patterns offer unique features that facilitate a PD analysis. They are a comparatively easy to acquire, language independent marker, since more or less well-defined standards for using them are established in the international scientific community. This information can be exploited to detect forms of plagiarism that cannot be detected with text-based approaches.[3]

The CbPD approach were unable to detect a single translated fragment ,the CbPD relies on citation information, it is unable to identify short paraphrased fragments. this approach is based on citation analysis and allows duplicate and plagiarism detection even if a document has been paraphrased or translated, since the relative position of citations often remain similar. Although this approach allows in many cases the detection of plagiarized work that could not be detected automatically with the currently used approaches, it should be considered as an extension rather than a substitute. Whereas the known text analysis methods can detect copied or, to a certain degree, modified passages, the proposed approach requires longer passages with at least two citations in order to create a digital fingerprint.

TABLE II
COMPARISON OF DETECTION RESULTS [4]

Plagiarism type	Text-based	Citation-based
Copy&paste	~ 70 % Good results even for short fragments	Unsuitable as short fragments cannot be detected
Disguised plagiarism	< 10 %	Depending on the fragments length ~ 30 %
Idea / structure plagiarism	0 %	Some cases could be identified
Translated plagiarism	< 5 %	~ 80 %. 13 out of 16 fragments could be identified.

C. SHAPE-BASED PD for flowchart

This method detects flow chart figure plagiarism based on shape-based image processing and multimedia retrieval.The method managed to retrieve flowcharts with ranked similarity according to different matching sets.There are many plagiarism detection system in which Most, if not all, discard the figures and charts before checking for plagiarism. Discarding the figures and charts results in look holes that people can take advantage. That means people can plagiarized figures and charts easily without the current plagiarism systems detecting it[6].

Therefore, there is a need to develop a system that will detect plagiarism in figures and charts there is a need to develop a system that will detect plagiarism in figures along with their contents. There is a need to develop a method for detecting figure plagiarism based on shape-based image processing where different types of figures can be considered for method for detecting figure plagiarism.

VI. SOFTWARE BASED PLAGIARISM DETECTION TOOLS

There are many software systems that suggest that they can reliably determine if a submitted text or an online document is plagiarized or not. Software can only hope to compare the syntax, on a character or word level, and determine the similarity between texts. There is some experimental work being done in the area of semantic recognition. But this only seems successful in the area of highly structured text such as program language code. [5]

A. PlagAware

Is an online-service used for textual plagiarism detection, which allows and offers some services for the user for example can search, find, analyze and trace plagiarism in the specified topic similar to the topics, PlagAware is a search engine, which is considered as the main element, which is strong in detecting typical contents of given texts. It uses the classical search engine for detecting and scanning plagiarism, and provide different types of report that help the user or the document owner to decide that is his document has been plagiarized or not.

B. PlagScan

PlagScan is online software used for textual plagiarism checker. PlagScan is often used by school and provides different types of account with different features. PlagScan use complex algorithms for checking and analysing uploaded document for plagiarism detection, based on up-to-date linguistic research. Unique signature extracted from the document's structure that is then compared with PlagScan database and millions of online documents.

C. CheckForPlagiarism.net

CheckForPlagiarism.net was developed by a team of professional academic people and became one of the best online plagiarism checkers that used to stop or prevention of online plagiarism and minimizes its effects on academic integrity.

In order to maximize the accuracy CheckForPlagiarism.net has used the some methods like document fingerprint and document source analysis to protect document against plagiarism.

D. iThenticate

iThenticate one of the application or services designed especially for the researchers, authors' publisher and other. It is designed to be used by institutions rather than personal, but lastly they provided a limit service for single plagiarism detection user like master and doctoral students and this allows them to check a single document of up to 25,000 words. So they can use this service to insure or to check their draft thesis whether containing correct citation and content originality

E. *PlagiarismDetection.org*

PlagiarismDetection.org: an online service provides high level of accuracy result in plagiarism detection. Mainly designed to help the teachers and student to maintain and to ensure or prevent and detect plagiarism against their academic documents. It provides quickly detect plagiarism with high level of accuracy

VII. CONCLUSIONS

This paper describes in brief the three different methods used for plagiarism detection. The Text-based PDS convince in detecting local forms of plagiarism, such as short passages of copied or only slightly paraphrased text. In contrast, they fail, to detect paraphrased and translated plagiarism. The citation-based approach is based on citation analysis and allows duplicate and plagiarism detection even if a document has been paraphrased or translated. Shape based plagiarism for flowchart presents a method for detecting flow chart figure plagiarism based on shape-based image processing but fails to detect plagiarism between different types of figures. Thus plagiarism detection system should not be based on single method but must be based on the combination of different plagiarism detection methods.

REFERENCES

[1] Hermann Maurer,Frank Kappe,Bilal Zaka" Plagiarism - A Survey"Institute for Information Systems and Computer Media Graz University of Technology, Austria, vol. 12, no. 8 2006

[2] Ahmed Hamza Osman, Naomie Salim1, and Albaraa Abuobieda," Survey of Text Plagiarism Detection" International University of Africa, Faculty of Computer Studies, Khartoum, Sudan Vol. 1, No. 1, June 2012

[3] Bela Gipp OvGU," Citation-based Plagiarism Detection – Idea, Implementation and Evaluation " Germany / UC Berkeley, California, USA 2010.

[4] Bela Gipp, Norman Meuschke, Joeran Beel" Comparative Evaluation of Text- and Citation-based Plagiarism Detection Approaches using GuttenPlag"

[5] Asim M. El Tahir Ali, Hussam M. Dahwa Abdulla," Overview and Comparison of Plagiarism Detection Tools "Department of Computer Science, Germany & UC Berkeley JCDL 2011.

[6] Senosy Arrish, Fadhil Noer Afif, Ahmadu Maidorawa and Naomie Salim" Shape-Based Plagiarism Detection for Flowchart Figures in Texts" Faculty of Computing, University Technology Malaysia, Skudai, Malaysia (IJCSIT) Vol 6, No 1, February 2014

