International Journal of Social Relevance & Concern
ISSN No: 2347-9698 Volume 2 Issue 10, October-2014

Detecting The Plagiarism For Text Documents On The World Wide Web

Durga Bhavani Dasari[1], Dr. Venu Gopala Rao. K[2]

[1]*Research Scholar, Dept of CSE, Jawaharlal Nehru Technological University, Hyderabad, India*

bhavani.dd@gmail.com

[2]*Professor, Dept of CSE, G. Narayanamma Institute of Technology and Science, Hyderabad, India*

Kvgrao1234@gmail.com

Abstract: *In the present age access, the world is growing with lots of easily accessible information available on almost every subject matter. The use of the freely available internet resource is causing easy copy and paste culture resulting in plagiarism in various research documents and academic reports. In such a scenario of the growing research and development publications, many techniques and methodologies have been developed for the plagiarism detection to evaluate the originality in the research documents both in regard to the web based as well as local repository based contents. This paper discusses various techniques and methods that detect and prevent plagiarism in articles, journals and scientific publications.*

Key words: Plagiarism Detection, Plagiarism prevention, Text mining, Web mining.

I. INTRODUCTION

Plagiarism by students, professors, industrialist or researcher is considered academic fraud. Plagiarism is defined in multiple ways like copying others original work without acknowledging the author or source. Original work is code, formulas, ideas, research, strategies, writing or other form. Punishment for plagiarism consists of suspension to termination along with loss of credibility. Therefore, detecting plagiarism is essential. Research paper selection is recurring activity for any conference or journal in academia. It is a multi-process task that begins with a call for papers. Fig. 1 shows research paper selection process. Call for paper is distributed to communities such as universities or research institutions. They are then assigned to experts for peer review. The review results are collected, and the papers arethen ranked based on the aggregation of the experts review results.

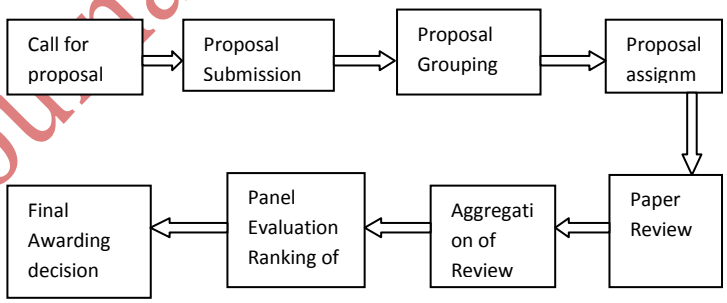

Figure 1*:* Research paper selection process

Expert reviewer may have inadequate knowledge inv research discipline.

Plagiarism detection software will help him to detect plagiarism quickly. Proposed system requires the database with existing research papers. When the call for papers (CFP) [1] is made from the end-user, the system accepts the research paper submitted by end-user. The system then finds the similarity between the paper submitted and existing research papers.

The proposed method aims to make manual process of checking plagiarism of research Papers computerised. The system allows an agency to ensure the ambiguity of the research Paper submitted by end-

user. It helps agency to find semantically similar research papers. Proposed method makes use of TF-IDF and LSI.

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 [2, 3,4].

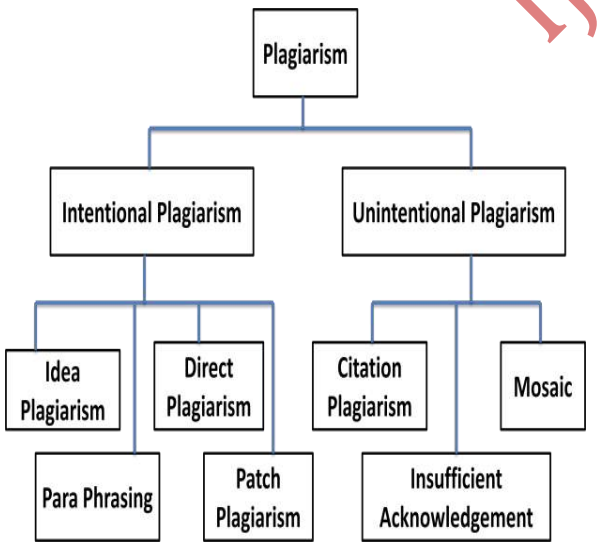

Figure 2: Classification of plagiarism

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 [5]. 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) [6]
Recall = (relevant items retrieved)/ (relevant items) [6]

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 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 [9].

3. 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 [10]. 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.

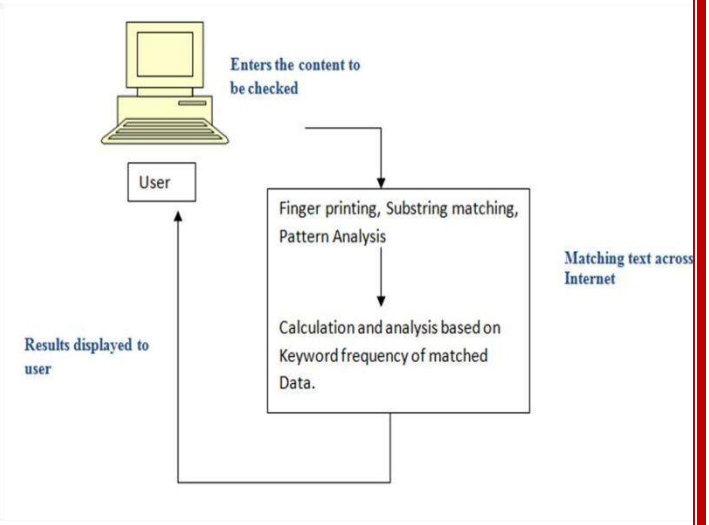

Figure 3: Plagiarism Detection Process

4. PLAGIARISM PREVENTION

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.

5. TEXT MINING

5.1 *Text document collection*:

The existing research papers are stored in the text format, within the database.

5.2 *Text document preprocessing*:

The contents of papers are usually non-structured. The pre-processing analyzes, extracts, and identifies the keywords in the full text of the papers and tokenizes them. Here, a further reduction in the vocabulary size is achieved, through the removal of frequently occurring words referred as stop-words, via-stop file. This is called as filtering phase of removal of stop-words.

5.3 *Text document encoding*

On filtering text documents they are converted into a feature vector. This step uses TF- IDF algorithm. Each token is assigned a weight, in terms of frequency (TF), taking into consideration a single research paper. IDF considers all the papers, scattered in the database and calculates the inverse frequency of the token appeared in all research papers. So, TF is a local weighting function, while IDF is a global weighting function.

6. 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 [11]:

$$\text{Purity}(\Omega, ¢) = \frac{1}{N} \sum_k \max_j | \omega_k \cap c_j |$$

Where $\Omega = \{\omega 1, \omega 2, \ldots, \omega K\}$ is the set of clusters and $¢ = \{c1, c2, \ldots, cJ\}$ 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 [11].

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.

7. CONCLUSION

This paper discusses about plagiarism detection, text mining and web mining and how to avoid the plagiarism on the web. Although there are the methods that detect plagiarism. In future should also shifted to multi lingual plagiarism detection as plagiarist could reuse the source from other language to their own. 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.

8. REFERENCES

[1]. Jian Ma, Wei Xu, Yong-hong Sun, Efraim Turban, Shouyang Wang, and OuLiu-An Ontology-Based Text-Mining Method to Cluster Papers for Research Project Selection, IEEE transactions on systems, man, and cybernetics—part a: systems and *humans, vol. 42, no. 3*, may 2012.

[2]. Loveleena Rajeev, Different types of plagiarism, July 18,2012.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] 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.

[6] 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.

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

[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] 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).

[11] 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.

