Academic Plagiarism Detection System

Prof. S.N.Zaware(Guide), Rajeev Pandey, Gaikwad Prashant, Abdulrahim Mulla
All India Shri Shivaji Memorial Society's Institute of Information Technology, Pune-411001

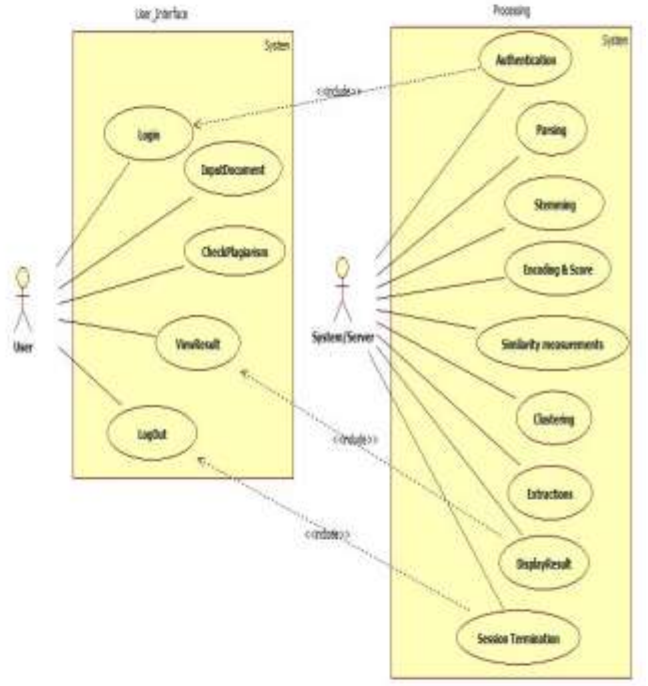
Abstract* - Existing plagiarism detectors have a huge limitation and one that isn't likely to go away any time soon. In fact, what they actually detect is sections of identical text only. So in our proposed system we intend to design an efficient, accurate as well as intelligent plagiarism detection software which will overcome and improve on the drawbacks of the existing plagiarism detection software. Here our main focus is on accuracy and efficiency of the detecting software. Here we are using hybrid algorithms like K-Means Clustering algorithm, TF-IDF, and combination of similarity measuring formulas like Cosine, Jaccard, Dice etc. In this paper, we will be comparing the input document on various parameters like its Grammar, Keywords, Sentences with the documents in the existing database. Based on the results obtained we will be display whether the input document is plagiarised or not.

***Keywords:* TF-IDF (term frequency, Inverse document frequency), K-Means clustering algorithm, Cosine, Jaccard, Dice .**

I. INTRODUCTION

Plagiarism comes from the Latin word *plagiarius*, which means abducting or kidnapping. Therefore, plagiarism is a fraudulent stealing or unlawful misappropriation of other author's text, language, thoughts, ideas, or expressions and then publishing them as one's own work, without giving proper credit or attribution to the original author's work. In fact plagiarism is not only an academic dishonesty or intellectual corruption but also a breach of journalistic principle and is subject to sanctions like penalties, suspension, and even removal. Though plagiarism is not a big crime but in academic world and industry, it is a serious moral offense and cases of plagiarism can sometimes amount to copyright infringement. Indeed, it is necessary to cite and acknowledge the sources even if those ideas are paraphrased and re-written with different words. Plagiarism is unethical and can hurt any person's or institution's reputation. So to overcome this problem we intend to design an efficient, accurate as well as intelligent plagiarism detection software. Here we are improving the plagiarism detection accuracy by using Hybrid algorithms i.e. TF-IDF score for finding important keywords , Combination of either Cosine & Jaccard or Cosine & Dice Depending on the best result obtained from this combination for Similarity measurement and K-means Clustering Algorithm[1] for finding the Relevant and Irrelevant documents in our existing database. In our Plagiarism detection Software we will be extracting the Features like Grammar, keywords and sentences of the input document and then we will compare all these features for plagiarism with the existing documents in our database. Our proposed System will certainly improve the accuracy and efficiency and will overcome the drawbacks of the Existing plagiarism Software.

II. SYSTEM MODEL



In our proposed system there are various steps of processing. So the first step is preprocessing which includes tokenization, Filtering, Stemming, finding Term frequency, finding Inverse Document frequency, and then calculating the TF-IDF score. Based on the TF-IDF score we will be finding important keywords. Now for comparison we need to find the relevant documents so that we can perform the similarity check for similarity between the two documents. So for that we are using K-means Clustering Algorithm. Here we will be making two clusters i.e. relevant cluster and irrelevant cluster. Then from our input document we will extract various features like Grammar, keywords, sentences using NLP parser and we will compare it the documents in the relevant cluster to find the percentage of similarity between two documents. Based on the output we will display the percentage of plagiarism detected. In this way we will design an efficient plagiarism detection software.

III. PREVIOUS WORK

Some of the present strategies for external plagiarism detection are supported string matching procedure [2], Vector space Model [3] and fingerprinting [4]. String matching procedure for plagiarism detection [2] aims to spot longest pairs of identical text strings. Suffix document models are principally used for the task. The strength of this procedure is that the detection accuracy with reference to the lexical overlaps. One downside of the rule is that the relative issue of police work disguised plagiarisms.

The tactic additionally needs vast machine efforts. Fingerprinting [4] is one among the foremost wide applied approaches for plagiarism detection. For a given suspicious document, its fingerprint is initial computed and also the trivialities are compared with a pre-computed index of fingerprints for all documents of a reference assortment. The inherent challenge of process is finding an exchange between document dimension and detection accuracy. The reduction in dimension for document illustration will cause to the knowledge loss that, in turn, affects system performance. the tactic needs standardization variety of parameters like the configuration strategy, chunk size (granularity of the fingerprint) or variety of trivialities (resolution of the fingerprint) [4]. Deciding the simplest parameter combination is powerfully smitten by the character and size of the document assortment additionally as on the quantity and also the styles of plagiarisms.

Citation-based plagiarism detection may be a pc assisted plagiarism detection approach which might be utilized in teachers. It doesn't consider the texts of the given documents however depend on the references given with a specific analysis paper. It initial identifies similar patterns within the citation sequences of 2 educational works .Subsequent non completely containing citations shared by each documents being compared is delineated by exploitation citation patterns. Citation patterns are known supported the factors of comparable order and proximity of citations inside the text. so as to quantify the pattern's degree of similarity, another factors, as an example absolutely the variety or relative fraction of shared citations within the pattern additionally because the likelihood that citations co-occur in a very document are thought-about. Stylometry applies some applied math strategies so as to see an author's distinctive style. It's principally used for characteristic the authorship attribution or intrinsic plagiarism detection.

IV. PROPOSED METHODOLOGY

A four stage processing algorithm is proposed here by combining clustering algorithm and similarity measurement formulas.

1) Pre-processing:

In this stage first we will take an input file which can be Pdf or any text file. This input file is sent to the system for performing various activities like tokenization, filtering, and stemming. In tokenization tokens are separated with the help string tokenizer class. These tokens are then sent for filtering the commonly used keywords like (is, or, the, for, of etc.). After filtering we these keywords are sent to stemmer. Here using porter

2) *TF-IDF:*

Example 1:
Let us assume, our Threshold value is 0.3
Consider a statement : "Civil Engineers build many kinds of modules"
Suppose we consider "civil"

(a) **TF value :**
Total numbers of tokens in this statement are 7

TF = Wi/Wt = 1/7 = 0.1429

(b) **IDF value :** Suppose there are 8 statements out of which 2

statements have the tokens "in them,Then

IDF = $\log(D_t=N_w) = \log(8=2) = 0:6020599$

(c) Now TF * IDF = 0.1429 * 0.60205 let us assume there are
Eight Statements out of which two statements have the
Tokens "in
Them, then Result = 0.08603

3. K-means Clustering Algorithm:

Here we will be making two clusters i.e. relevant cluster and irrelevant cluster using the threshold value which will be calculated dynamically.

4. Feature Extraction and Comparison:

Here we will extract the features like Grammar, Keywords, and Sentences of the input document using the NLP parser and then we will compare these features with the existing document in the database. Based on the final result of comparison we will display whether the input document is plagiarized or not.

V. CONCLUSION

This project aims at minimizing fraudulent stealing of another writer's ideas or expressions by detecting whether the Paper which is to be published is Plagiarized or not. So in our system we intend to design an efficient, accurate as well as intelligent plagiarism detection software. Here we are improving the plagiarism detection accuracy by using Hybrid algorithms i.e. TF-IDF score for finding important keywords , Combination of either Cosine Jaccard or Cosine Dice Depending on the best result obtained from this combination for Similarity measurement and K-means Clustering Algorithm for finding the Relevant and Irrelevant documents in our existing database. Our proposed System will certainly improve the accuracy and efficiency and will overcome the drawbacks of the Existing plagiarism Software.

REFERENCES

[1] M.Sathya , J.Jayanthi, N. Basker, “Link Based K-Means Clustering Algorithm for Information Retrieval"in IEEE-International Conference on Recent Trends in information Technology, ICR-TIT 2011.

[2] Kim, Plagiarism detection using the levenshtein distance and smithwaterman algorithm, in Proceedings of the2008 3rdInternational Conference on Innovative Computing Information and Control, ser. ICICIC 08. Washington, DC,USA: IEEE Computer Society, 2008, pp.

[3] J. Kasprzak and M.Brandejs, Improving the reliability of the plagiarism detection system-lab report for pan at clef 2010, in Notebook Papers of CLEF 2010 LABs and Workshops, 2010. Name of Auhtors, “Title of the research”, Citation Details, year.

[4] T. C. Hoad and J. Zobel, Methods for identifying versioned and plagiarized documents, J. Am. Soc. Inf. Sci. Technol., vol. 54, no. 3, pp.203215, Feb. 2003.