International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 4 Issue: 2 175 - 178

Plagiarism detection for document

Varun Shukla
University of Mumbai,
Information Technology Dept,
K.J.S.I.E.I.T
Mumbai, India
varun.shukla@somaiya.edu

Farhana Khan
University of Mumbai,
Information Technology Dept,
K.J.S.I.E.I.T
Mumbai, India
farhana.khan@somaiya.edu

Komal Mody
University of Mumbai,
Information Technology Dept,
K.J.S.I.E.I.T
Mumbai, India
komal.mody@somaiya.edu

Prof Sarita Rathod
University of Mumbai,
Faculty of Information Tech.Dept,
K.J.S.I.E.I.T.,
Mumbai, India
sarita.r@somaiya.edu

Abstract-Our project aims to provide plagiarism based on semantic detection and natural language processing technique. Plagiarism detection for document is very effective technique, as nowadays students are mainly dependent on Internet. . The wide use and availability of electronic resources makes it easy for students, authors and even academic people to access and use any piece of information and embed it into his/ her own work without proper citation. Our project help authors, writers etc. to secure their files and make their files safe. It helps the user to upload the file easily and detect plagiarism more efficiently. It gives the more accurate results. This web application will help the users to upload the files and check for the plagiarism more easily and securely.

Keywords*: Natural language processing, Semantic detection, Plagiarism Detection.*

*****

I. INTRODUCTION

Plagiarism is a very common phenomenon now days and it is center of discussion at various educational events. Plagiarism is defined as the practice of claiming or implying original authorship of (or incorporating material from) someone else's written or creative work, in whole or in part, into one's own without adequate acknowledgement. This seems to be a temporary solution for this problem because eventually we are trying to control the possible cases of plagiarism. Most of the research in this field has till now concentrated in the type of text extraction which in turn can be sentence or word Chunking, no one has yet bi-furcated the extraction into sub parts such as rearrangement to explore the performance issues.

There are two different approaches to automatic plagiarism detection:-

1. External or extrinsic plagiarism detection is based on detecting the similarity of the source document with the documents present in the reference text dataset.

2. Intrinsic plagiarism detection approach is based on detecting the plagiarism that exists in a suspicious text itself without having a reference text dataset.

The plagiarism detection approach in this work is based on the problem of detecting the plagiarized documents by making use of an existing reference text dataset. Hence this proposed work is an extrinsic monolingual plagiarism detection approach which identifies whether the suspected documents are plagiarized versions of a given source document.

Advantage of Proposed System:-

- It saves the time and the user spending cost.
- It saves and gives security to the users.
- It protects user's data from being copying.
- It is allow the user to dynamically save a new file through the web

The rest of the paper is organized as follows: Section II gives a brief idea of the problem that exists and idea how it will be solved by our proposed system. Section III provides comparisons of various existing system. Finally, the conclusion and the major contribution to this paper are discussed in the remaining sections.

II. PROBLEM DEFINITION:

Plagiarism is a serious phenomenon that is spreading widely in all levels of academia as well as in all sorts of intellectual work. Plagiarism or Digital Plagiarism as mentioned by Butakov.&Scherbinin. (2009), has many forms and therefore many definitions, according to Alzahrani et al. (2009) it may include copying part or all of the text and use it without referring it to the original composer, rephrasing the text by changing the words used but expressing the same ideas from others work, translating others work from language to language without referencing the right references and source code cloning which is done by using any piece of programming source code developed by other programmers

and reuse without proper citation or even permission to. Education expert Ryan. (2007) has discussed how serious the problem of plagiarism is, and how it seriously produce poor academic performance; because simply keeping the quality of learning will not be guaranteed if we couldn't maintain the quality of all learning processes including assessment process which is one of the most important processes that is hardly affected by plagiarism[7,8,9]. The solution to plagiarism can be maintained by following two approaches, plagiarism prevention and plagiarism detection. Alzahrani et al. (2009) explained that Prevention will be achieved by providing the ability of detecting the unoriginal content of student assignments or any kind of work thus affecting judging and assessing that work. Our project intents to provide a place for plagiarism detection where the user can easily detect and upload their files for ahead reference. According to Total Quality Management principles, to insure quality of learning process all phases of learning - including assessment - should comply to quality standards, and to achieve this specially when dealing with challenges of Online Education, assessment should be controlled by automated tools for plagiarism checking and detection.

III. LITERATURE REVIEW

1. The IEEE paper "Automatic Cross-Language Plagiarism Detection "published by authors Angel ANGUITA, Alejandra BEGHELLI and Werner CREIXELL uses the cross-language plagiarism in electronic documents.

2. The IEEE paper "Web based Cross Language Semantic Plagiarism Detection " by authors Chow Kok Kent and Naomie Salim uses the cross language and cultural border and with different types of translation tools, cross language plagiarism is bound to rise.

3. The International Journal paper "Automated Plagiarism Detection System for Malayalam Text Documents" by authors Sindhu. L, Bindu Baby Thomas and Sumam Mary Idicula uses plagiarism detection tool for plagiarism detection in Malayalam documents is presented.

4. The International Journal paper "CHECK: A Document Plagiarism Detection System" by author Bela Gipp compares the occurrences of citations in order to identify similarities.

5. The British Journal of nursing paper "Step-by-step guide to critiquing research. Part 2: qualitative research" by Frances Ryan, Michael Coughlan, Patricia Cronin researches on quantitative study.

6. The Journal paper "Statement-based fuzzy-set IR versus fingerprints matching for plagiarism detection in Arabic documents" by Alzahrani SM, Salim N compare the documents against the intra corpus collection, which probably contains the previous assignments. Moreover, APD tool searches the web to give similar resources as well. An automatic report will be generated that contains highlighted plagiarized parts and a list of similar resources ranked from highest to lowest

7. The IEEE International Conference paper "On the number of search queries required for Internet plagiarism detection" by Sergey Butakov In the digital era, with all the information now accessible at students' fingertips, plagiarism detection services (PDS) have become a must-have part of LMS. In most such systems, to compare a submitted work with possible sources on the Internet, the university transfers the student's submission to a third-party service. Such an approach is often criticized by students, who regard this process as a violation of copyright law. To address this issue, this paper outlines an improved approach for PDS development that should allow universities to avoid such criticism. The major proposed alteration of the mainstream architecture is to move document preprocessing and search result clarification from the third-party system back to the university system. The proposed architecture changes would allow schools to submit only limited information to the third party and avoid criticism about intellectual property violation.

The proposed system is compared with the following existing systems as shown in table 1,

Table 1- System Comparison

Existing System	Proposed System
Language wise detection	Overall detection
Paragraph wise detection	Line by line wise detection
Intersic detection	Extensic detection
Poor report generation	Better report generation
Poor Performance	Better Performance
Similar sounding detection is not done	Similar sounding words detection
Active passive voice detection is not done	Active passive voice detection

IV. METHODOLOGY

1. In the proposed system we will detect the plagiarism sentences by sentences. The system must also detect similar sound text so that no one can replace text through similar sound text. We are also trying to detect active and passive voice plagiarism.

2. It works with module by module ,each module independent of each other working together and a report is generated that the document is plagiarized or not.

3. We are designing the front end with HTML 5.
4. We are programming back end with .NET Framework.

5. We are using Microsoft SQL database.

The project aim is to reducing the amount of time spent comparing texts, making comparison between large numbers of multiple texts feasible and finding possible source texts from electronic resources available to the system. The systems must minimize the number of incorrectly classed as plagiarized and those incorrectly classed as non-plagiarized and maximize the number of true positives. The system must also detect similar sound text so that no one can replace text through similar sound text. We are also trying to detect active and passive voice plagiarism.
Example: - The Words that sounds similar like “Smith” and “Smythe” can be detected using our project. The words with similar meaning will only be detected.

Also active passive voices like “He is watching movie” and “movie is watched by him” will also be detected using this project.

Figure 1 shows the following steps Steps:-

1. User uploads the file.
2. The file is in the format of doc, text only.
3. If multiple users upload the file together one by one file is sent for upload purpose.
4. File is saved in the temporary folder.
5. In this step, “Plagiarism algorithm” is executed; file of the user is checked with each and every file in the database. Each file is checked sentence by sentence, whether it is copied or not.
6. If any of the line or phase is copied then the file will not be uploaded and the file will be deleted from the temporary folder and user get the message that “file has been plagiarized”.
7. If all the sentences are different, then the file of the user is successfully uploaded and entered to the database.

Figure 2 shows the System Block diagram and components are:-

Input: - It is the input of the user. User will upload the file which he/she wants to upload in the webpage and go for plagiarism detection.

Detector: -It is there, if the file which is being uploaded by the user is doc file or not. If Doc file is not there then it will not go ahead for plagiarism.

Checker: -Here plagiarism algorithm is implemented and file is checked with the entire files from the database.

Database: -All the files are saved in the database. Whenever a file is uploaded by the user it is checked for the plagiarism with the entire files from the database only.

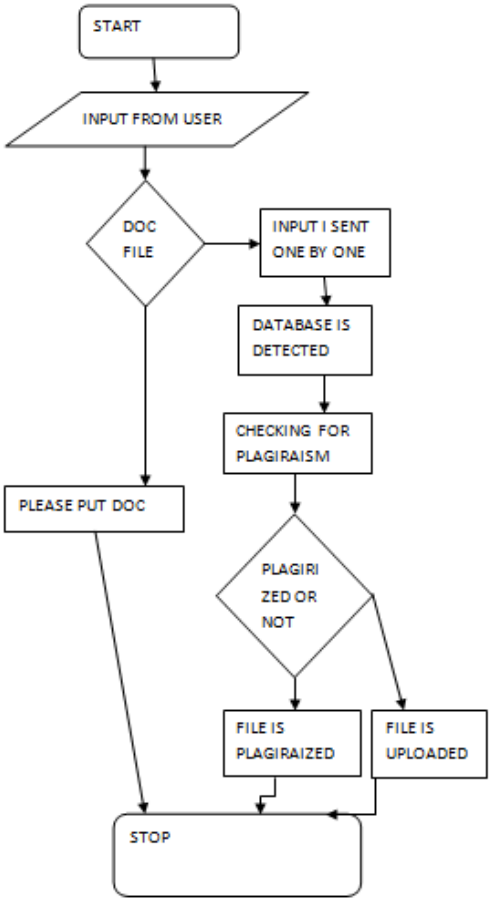

Figure 1: System Architecture

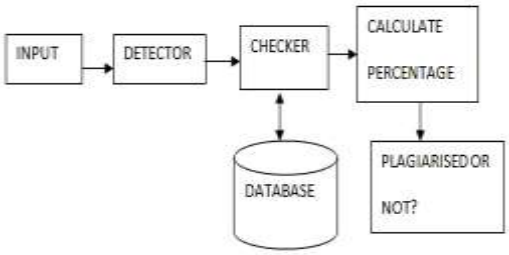

Figure 2: System Block Diagram

V. CONCLUSION

In this paper we presented our web based cross language semantic plagiarism detection approach. Nowadays, semantic plagiarism has becomes a major issue and concern especially in the academic works. Most of the current plagiarism detection tools are not stable against both of these plagiarism cases. With our experiment, the result shows that our proposed techniques achieve significant improvement. We are also detecting similar sounding and active passive voice detection.

REFERENCE

[1] Angel ANGUITA, Alejandra BEGHELLI and Werner CREIXELL "Automatic Cross-Language Plagiarism Detection" IEEE paper Electronic Eng. Dept., Universidad Tecnica Federico Santa Maria Valparaiso, Chile and Electronic Eng. Dept., Universidad Tecnica Federico Santa Maria, Valparaiso, Chile CSIS, University of Tokio (Visiting researcher), Japan 2011.

[2] Chow Kok Kent and Naomie Salim "Web based Cross Language Semantic Plagiarism Detection" Universiti Teknologi Malaysia Johor, Malaysia and Universiti Teknologi Malaysia Johor, Malaysia , IEEE paper 2011.

[3] Sindhu. L, Bindu Baby Thomas and Sumam Mary Idicula "Automated Plagiarism Detection System for Malayalam Text Documents" International Journal of Computer Applications (0975 – 8887) Volume 106 – No. 15, November 2014.

[4] Amandeep Dhir , Gaurav Arora, Anuj Arora "ARCHITECTURAL DESIGNING AND ANALYSIS OF NATURAL LANGUAGE PLAGIARISM DETECTION MECHANISM" Journal of Theoretical and Applied Information Technology , 2005 – 2008.

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

[6] Antonio Si, Hong Va Leong, Rynson W. H. Lau "CHECK: A Document Plagiarism Detection System" In Proceedings of ACM Symposium for Applied Computing, pp. 70-77, Feb. 1997.

[7] Alzahrani SM, Salim N. "Statement-based fuzzy-set IR versus fingerprints matching for plagiarism detection in Arabic documents"In Proc. of the 5th Postgraduate Annual Research Seminar, Malaysia; 2009.

[8] Butakov, S., & Shcherbinin, V. (2009)" On the number of search queries required for Internet plagiarism detection" In Proceedings of the 2009 Ninth IEEE International Conference on Advanced Learning Technologies (pp. 482–483). Washington, DC: ICALT, IEEE Computer Society.

[9] Frances Ryan, Michael Coughlan, Patricia Cronin "Step-by-step guide to critiquing research. Part 2: qualitative research" British Journal of Nursing, 2007, Vol 16, No 12.