An Approach for Implementing Web-Based Tool for Plagiarism Detection

Jayanti Khatri¹, Vishal Mohan²
¹Department of Computer Engineering, Poornima University, Jaipur, INDIA
²Department of IT, Dronacharya College of Engineering Gurgaon, INDIA

ABSTRACT
Plagiarism detection is the process of detecting instances of plagiarism within a work or document. The widely distributed use of computers and internet has made it easier to copy the work of others without giving the credit of originality of the owner’s document. The need for detecting plagiarism still exists and automated tools can provide valuable assistance in this task. Most such tools developed so far have focused on analyzing the content of the final work submitted by student.

In contrast this paper describes a web-based anti plagiarism approach to detect plagiarism at the academic level. This system will detect the plagiarism in student’s assignments, seminar reports and project reports and it will also check for teacher’s research papers, thesis submitted to the department in the various fields of research done at P.G level.

Keywords--- academic integrity, Plagiarism detection, similarity,

I. INTRODUCTION

1.1 DEFINING PLAGIARISM
Plagiarism, as defined in the 1995 Random House Compact Unabridged Dictionary, is the “use or close imitation of the language and thoughts of another author and the representation of them as one's own original work [1].” With the widespread of internet and the vast amount of text available electronically have made people easier to plagiarize the work of other’s and the investigations done to detect plagiarism are very few.

Plagiarism is “ representing the intellectual property of others or the work of a third party as ones own work, totally or in part”[1]

The things that immediately come to mind as description of plagiarism are:

a) Turning in someone else’s work as your own.

b) Copying words or ideas from someone else without giving credit.

c) Failing to put a quotation in quotation marks.

d) Giving incorrect information about the source of quotation.

e) Changing words but copying the sentence structure of a source without giving credit.

f) Copying so many words or ideas from a source that it makes up the majority of your work, whether you give credit or not. [2]

1.2 CLASSIFICATION OF PLAGIARISM
Plagiarism can be broadly classified into categories that include:

• Accidental: Due to lack of plagiarism knowledge, and understanding of citation or referencing style being practiced at an institute.

• Unintentional: The vastness of available information influences thoughts and the same ideas may come out via spoken or written expressions as one's own.

• Intentional: A deliberate act of copying complete or part of some one else’s work without giving proper credit to original creator.

1.3 PLAGIARISM DETECTION
Plagiarism detection is defined as detecting instances of plagiarism within a work or document. This paper discusses an approach for the implementation of a plagiarism detection tool which will detect plagiarism at institute or academic level. The approach uses the similarity search method for detecting plagiarism. The users who are the students are allowed to upload their documents. This document will be permanently saved on the system database. When a search for similar document is requested this database consisting of other users documents can also be searched to detect plagiarism.

II. BACKGROUND
As the advent of technology is growing day by day there is an alarming increase in plagiarism in academic institutions. This gradual increase of plagiarism in academic institutions is putting a negative impact on the academic integrity of the institution. It is becoming very difficult to translate values, into action and a stringent action is badly needed to promote academic integrity in campuses. Higher education and society benefit when academic institutions have a very bright and striking standards of integrity that provide foundation for a resonant academic life, promote scientific progress and prepare students to be a responsible citizen.

The Centre of Academic Integrity (CAI) defines academic integrity as a commitment even in the face of adversity to five fundamental values: honesty, truth, fairness, respect and responsibility. From these values flow principles of behavior that enable academic communities to translate ideas into action.

Honesty: The quest for truth and knowledge requires intellectual and personal honesty in learning, teaching, research and service.

Trust: Academic institutions must foster a climate of mutual trust in order to stimulate the free exchange of ideas.

Fairness: All interactions among students, faculty and administrators should be grounded in clear standards, practices and procedures.

Respect: Learning is acknowledged as a participatory process, and a wide range of opinions and ideas is respected.

Responsibility: A thriving community demands personal accountability on the part of all members and depends upon action in the face of wrongdoing.

Acts of academic dishonesty like plagiarism compromise these core values and undermine the process by which knowledge is created, shared and evaluated.

The main objectives of proposing this approach are as follows:
1. To implement the five fundamental values of academic integrity i.e. Honesty, Trust, Fairness, Respect and Responsibility and lay a foundation for responsibility towards society in students lives after graduation.
2. To develop an effective and rigid web based plagiarism detection system which will help in to stop plagiarism as it is not so much implemented and more not so given much importance in society.
3. To implement this system at academic level and to bring a halt to plagiarism this is very much profoundly found in academic institutions.
4. To detect and penalize the work done by student who complete their work (assignment, seminar and project reports) by implementing plagiarism and devalue the work of students who have done original work.

III. RESEARCH PROCESS

The proposed approach compares suspect document to a large collection of other documents and makes an attempt to match parts of the suspect document to parts of those in the collection. It also makes use of the Google Search API which can be used for certain keywords or key sentences from a suspected document on the World Wide Web. This method is highly effective for checking students assignments, reports (project/seminar).

Fig 1. Working of “Plagiasi”

The steps involved in developing this tool are as follows:
1. Input Phase
2. Validation and storage Phase
3. Fragmentation Phase
4. Matching Phase
5. Result Phase
6. Output Phase

1. Input Phase:

This the first phase of the system in which the user will login into the system using his/her login name and password. The system will authenticate the user by checking the user name and password.

When the system will authenticate the user then the control transfers the user to the main page where the user will upload his/her documents for plagiarism detection.

These two phases are directly accessing the database.

2. Validation and Storage Phase:
The second phase deals with the validation and the storage operation. In this phase the uploaded document by the user will be validated whether the document is in the acceptable format (.txt, .doc) or not by this system. If the uploaded document is in the acceptable format the system will pop the message when the document will be uploaded successfully.

If the document is not in the acceptable format then system will throw an error message “File format not supported” and give the details of the acceptable file format details to the user.

3. Fragmentation Phase:
The third phase deals with the uploaded document. In this phase the document is split up into the small fragments (chunks) in the Strings, phrases (according to the basic concept of search engine).

String: The document split up into the small fragments (chunks) which are converted into Strings and Phrases. The string is collection of character separated by space.
Ex. Academic Integrity

In the above example the fragmented will be generated as [Academic] as first string and [Integrity] as a second string.

Phrase: In the uploaded document when the fragments (chunks) are created the String and phrases will be generated. The String which are in double quotes (“”) will be treated as the phrase and will be matched as it is with database.
Ex. “Academic Integrity”

In the above example the whole string will be treated as the phrase [Academic Integrity] and will be matched as it with local, distributed and global database.

4. Comparison Phase:
In the Previous phase all the generated fragments (chunks) will be stored locally in the database. All these fragments (chunks) then will be matched with the Local database (local drives), Distributed database (LAN), and Global database (www) sequentially.

For comparing two documents, the following steps are computed:
- First each document is partitioned into tokens.
- Part- of Speech disambiguation.
- Find the most appropriate sense for every word in a document
- Finally, compute the similarity of the sentences based on the similarity of the pairs of words.
This phase is very important phase in terms of the document comparison and database usage.

5. Result Phase:
When all the comparison of fragments (chunks) is done with the Local database (Local drive), Distributed database (LAN), and Global database (www) then the system will display the result in form of URL detected through Global database, the path of the file saved locally (Local drive) and the different name of the systems connected in the network i.e. on the distributed database.

6. Output Phase:
This is the last phase of the system in which all the output is collected through the plagiarism detection system and the report will be generated according to the comparison of the fragments (chunks) to the Local, Distributed and Global database and will be generated on the web page.

The result will be in the form of the percentage of true copy of the document. It will also display the URL i.e. the source from which the user has prepared his document.

IV. RESULTS

The outcome or the expected final report that will be generated will be in the following format.

![Fig 2. Expected Result](image)

V. CONCLUSION

In the age of information technology plagiarism has become more actual and has become a serious issue for the society and academic institutions. Plagiarism prevention methods which need societal change of attitude against plagiarism are the best means to fight against plagiarism.

Our approach for plagiarism detection introduces a method for online detection of plagiarism and tries to bring a halt to the cut, copy and paste culture which is going on a vertical extent in academic institutions. Thus the implementation of approach will try to stimulate the intellectual growth and development among students and will make understand the five fundamental values of academic integrity upon which knowledge is created, shared and evaluated.
REFERENCES

[1] German Wikipedia:
   en.wikipedia.org/wiki/German_language
[3] The Fundamental Values of Academic Integrity:
   www.academicintegrity.org/fundamental_values/pdf
[6] Plagiarism ACM document: An Anti Plagiarism Editor for Software Development Courses by Peter Vamplew and Julian Dermoudy, School of Computing University of Tasmania