DETECTING PLAGIARISM AND TEXT SUMMARIZATION

[1]Shima Gopi, [2]Priya.C.S
[1]Student, [2]Lecturer
[1]Computer Science,
[1St].Joseph's College, Thrissur, India

Abstract : ***Text summarization helps to get the compressed form of original document. It is mainly categorized by two type. They are extractive and abstractive summarization. But these two techniques has lot of limitations. Text summarization is also called document summarization which provide more information in less space. Author implementing plagiarism on Text summarized data to identify Unjust appropriation, theft and publication of another persons creation including language, thoughts ,idea or expression and representation it as one's own.***

Index Terms – Text Summarization,Plagiarism,Copyright,Rough set,Histogram.

I. INTRODUCTION

The system implements the plagiarism on text summarized data which can widely used in many of the applications in todays world. Text Summarization aims to generate concise and compressed form of original documents. Author implementing plagiarism on Text summarized data to identify Unjust appropriation, theft and publication of another persons creation including language, thoughts ,idea or expression and representation it as one's own. Plagiarism so defined could refer to copying paragraphs from the Internet without citation, slightly modifying parts of a fellow student's for an assignment, paying a third party to do the work, and so on. Text summarization aims to generate compact form of the original text.The techniques used for summarizing text may be categorized as extractive summarization and abstractive summarization. Author used extractive summarization for summary.extractive summarization performed by using reduct[2].three types of matrixes using for find the reduct set[2]. discerniblity matrix ,indiscernibilty matrix and equal to one matrix are the matrixs.

II.EXISTING SYSTEM

The Existing system does not implement the plagiarism technique on text summarized data.Hence we cannot apply the plagiarism on summarized data.

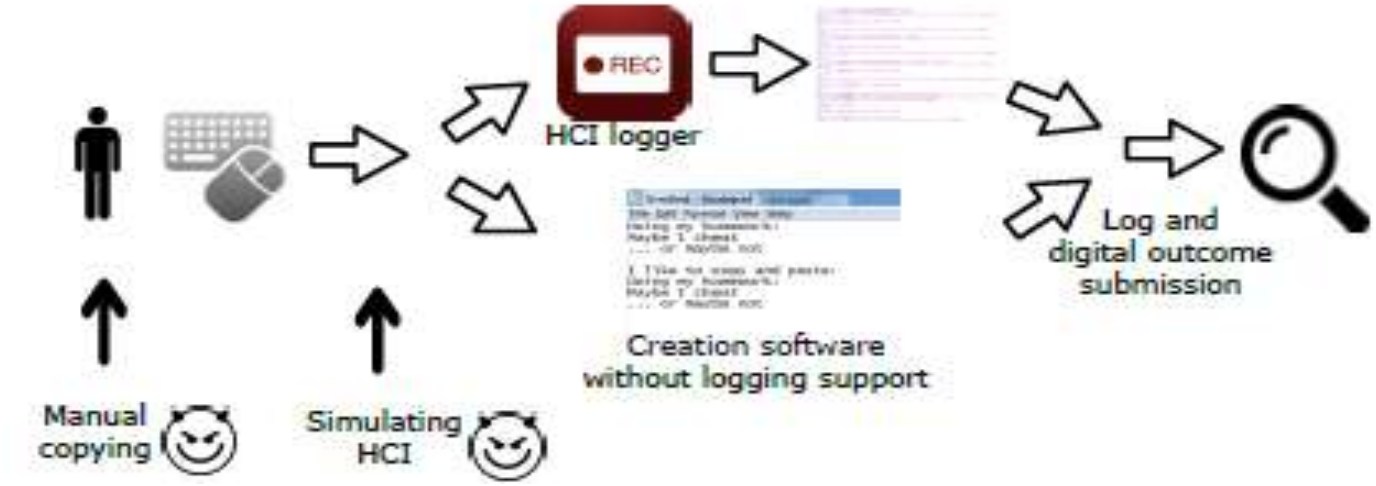

figure 1; log creation by tracking human computer interaction directly

The setup in Figure 1 for tracking HCI is more general and more widely applicable than is reliance on logs from the creation software since not all applications support logging.In determining whether a text has been plagiarized or not, text analysis can be carried out. Statistics is used to decide whether a text has indeed been plagiarized based on the frequency counts on the words and sentences.

III.PROPOSED SYSTEM

This paper focused on extractive techniques. We address the issue of extraction of major sentences by means of rough sets[2] notion of reduct[2].Also provides the plagiarism detection technique.

IV.METHODOLOGY

The important phases in this proposed method are,

File uploading: Uploading different types of files.
Text summarization: Summarizes the text in convenient form.
Plagiarism: Checks plagiarism of that uploaded document.
Combined result: Combined result of summary of document and plagiarism analysis.

Algorithm 1 Text Summarization
Input: Text Document.
Output: Summary sentences based on reduct
1: Creating information table from a text document.
2: Generate matrices.
3: Reduct Construction [Algorithm 2]

Algorithm 2 Reduct Construction [27]
Input: A matrix *M* is either *MDIS*, *MIND* or *MEO*.
Output: Reduct as union of all elements in *M*
1: for $k = 2$ to n do
2: for $l = 1$ to $n - 1$ do
3: if $M(k, l) = \emptyset$ then
4: for $M(k', l') \in$ B and $M(k', l') = \emptyset$ do
5: if $M(k', l') \subset M(k, l)$ then
6: $M(k, l) = M(k', l')$
7: end if
8: end for
9: end if
10: Split $M(k, l)$ into two
11: Pick a from $M(k, l)$, $(a \in At)$;
12: $A = M(k, l) - \{a\}$
13: $M(k, l) = \{a\}$
14: Simplify all elements in B that are non empty
15: for $\forall M(k', l') \in$ B where $M(k', l') = \emptyset$ do
16: if $a \in M(k', l')$ then
17: $M(k', l') = \{a\}$
18: else
19: $M(k', l') = M(k', l') - A$
20: end if

Text summarization performed by using rough set.in rough set Reduct[2] is a attribute subset that contain the same information as the original set of attributes[2].reduct representing text in the form of an information table.The information table contains a finite set of objects which are defined by a finite set of attributes.Information table S is mathathatically represented as

$S=(U,At,\{ Va|a \in At\}, \{Ia|a \in At\})$.

Where U denotes a set of objects. *At* denotes a finite set of attributes. *Va* is the domain for an attribute $a \in At$ and *Ia* is an information function which maps an object $x \in U$ to a value of *Va*.

Author consider three types of matrices,namely,discernibility matrix,indiscerniblity matrix and equal to one matrix. discernibilty matrix identified the special objects called exceptions.Removing exceptions from the original data author can be obtain more concise text and learn some information.

*Discernibility matrix:a discernibility matrix MDiss is a U*u matrix where a particular entry of the matrix is defined as,MDiss(x,y)={a∈At | Ia(x)= Ia(y)}*

This means that MDiss(x,y)will contain all those attributes based on which we can discern,distinguish or differentiate between objects x and y.

*Indiscernibilty matrix: an indicernibilty matrix MInd is a U*U matrix where a particular entry of the matrix is defined as Mind(x,y)={a∈* At | *Ia* (x) = Ia(y)}.

This suggests that Mind(x,y) will contain those attributes based on which are unable to discern between objects x and y.

Equal to one matrix: an equal to one matrix MEo is a U*U matrix where a particular entry of the matrix is defined as

,$MEo(x,y)=\{a \in At|Ia(x) = Ia(y) = 1\}$.

An important is how to obtain the minimum matrices by simplyfing the original matrices.for solving this problem by using element absorbtion method and element delection method.

Element absorbtion menthod: it absorb another element $M(x, y)$ if,$\phi \neq M(x', y') \subset M(x,y)$.

This means that the values of M(x,y) will be replaced by M(X',Y').This operation is based on the rationale that the attributes in M(x',y') are enough to distinguish that object pairs.that is (x',y') and (x,y).after this operation ,no element in M(x,y)is a proper subset of any other element in M(x,y).

Element Deletion: The element deletion operation deletes *a* from all the elements of the matrix if,

$\forall(M(x, y) \neq \phi)[(M(x, y) - \{a\}) \neq \phi]$

The attribute deletion operation will delete attributes as long as we still able to discern object pair (x, y). The reduct construction algorithm will repeatidlt apply element absorption[2] and element deletion[2] operations until minimum matrix is obtained. Several special techniques address source code plagiarism.For example, [4] introduces plagiarism detection of source code by using execution traces of the final program, that is,method calls and key variables. An API-based control flow graph used in [6] to detect plagiarism represents a program as nodes (statements), with an edge between two nodes if

there is a transition in the graph between the nodes. API-based control flow graphs merge several statements to an API call, which extracts features and then builds a classifier to detect plagiarism. Focusing on detecting plagiarism for algorithms, [3] relies on the fact that some runtime values are necessary for all implementations of an algorithm, so any modification of the algorithm will also contain these variables.Plagiarism detection is performed by two areas.

A. Data preparation

Data cleaning and transaformation involved in data preparation.in data cleaning ,the entire logs as well as cleaning content of individual logs.these logs removed since a cheater is not likely to copy an incomplete assignment.we computed a histogram[1] for each log that captures the frequency of each command type.

B. 'copy and modifi' detection

The goal of copy and modify detection is to determine if one log is an altered version of another by computing the similarity between two logs.we compute the correlation for randomly chosen subsets of commands.for a pair of logs to be similar it suffices that the correlation for just one of the subsets is similar.command types must be chosen for correlation computation separatey for each pair of logs,rather than choosing just one subset of command types for all logs.

V.IMPLEMENTATION

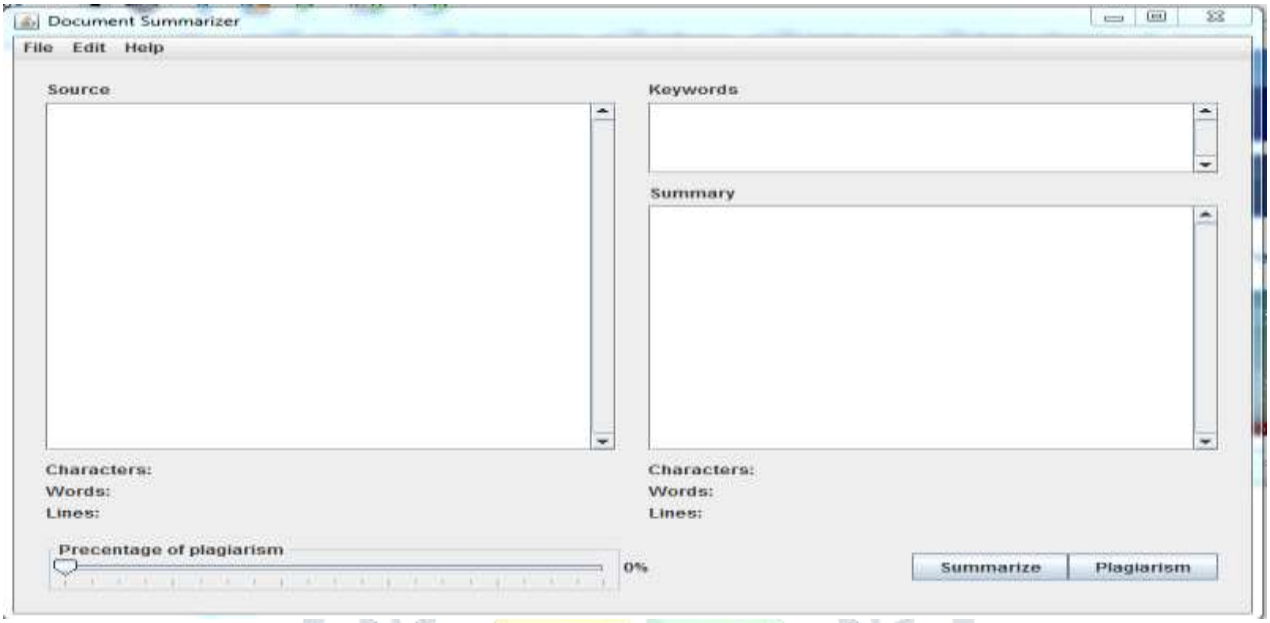

figure:2, File uploading: Uploading different types of files.

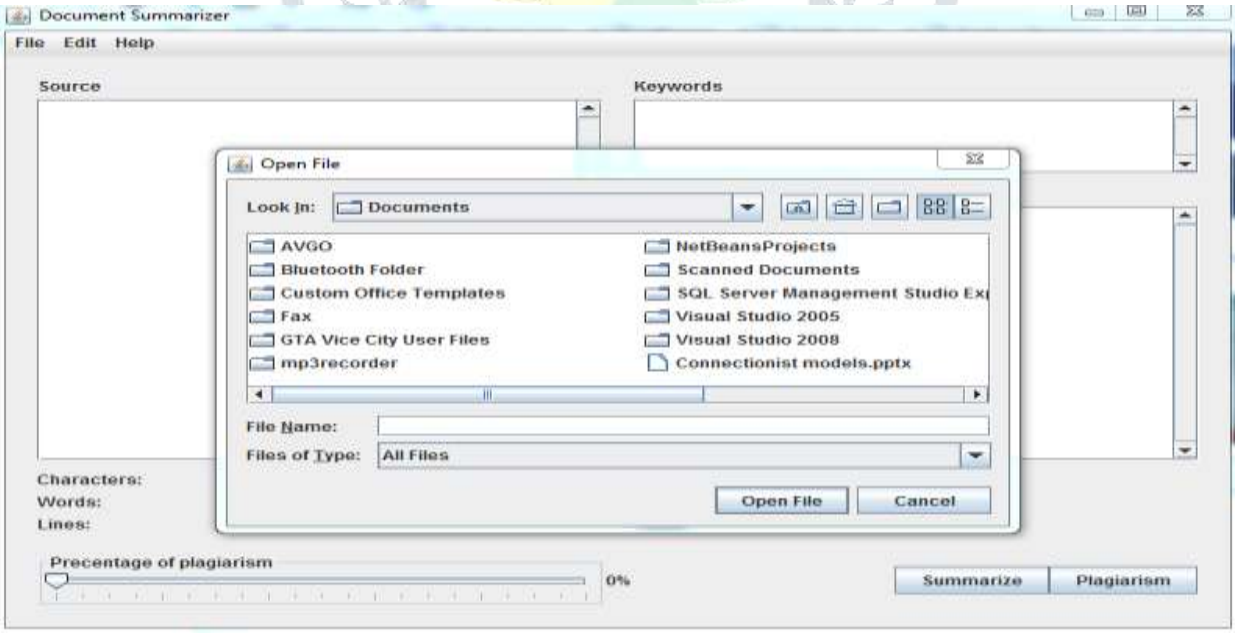

figure:3, Text summarization: Summarizes the text in convenient form.

figure:4, Plagiarism: Checks plagiarism of that uploaded document.

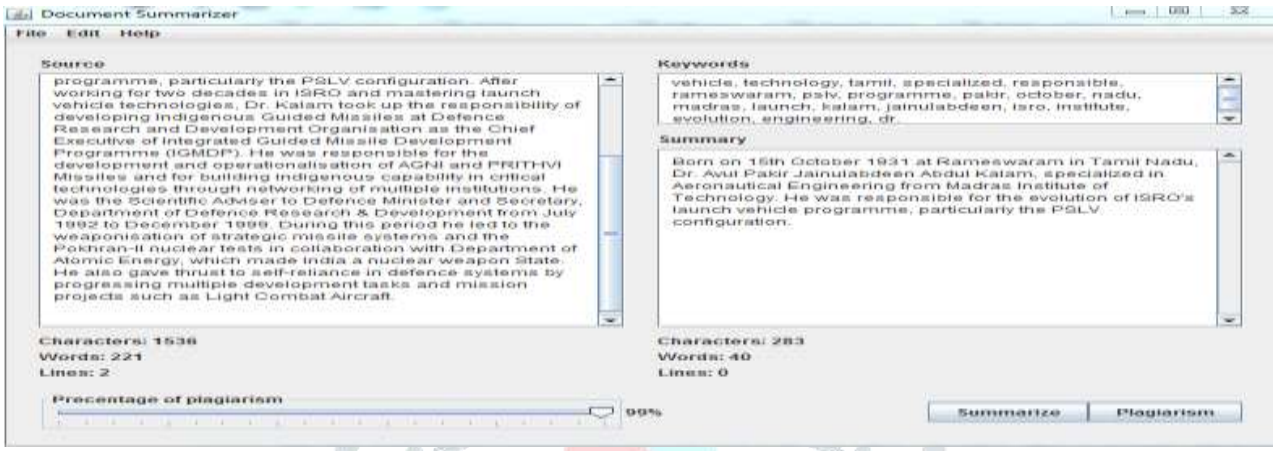

figure:5, Combined result: Combined result of summary of document and plagiarism analysis.

VI. CONCLUSION

Proposed system solves majority of disadvantages of existing system. Combined text summarization and plagiarism increases efficiency..Finding out the information related to the need of a user among large number of document is a problem that has come with the growth of text based resources.in order to solve this problem,summarization methods are proposed and evaluated.The summary contains all necessary information with out redundant data.One of the possible future work is display plagiarism with name of article ,size of copy,original author of article,model of copying. .

VII.ACKNOWLEDGEMENT

First of all, I am grateful to The Almighty God for establishing me to complete this project. I am especially thankful to my guide, Ms.Priya.C.S, and all other faculty members from the Department of Computer Science and my friends, for giving me their sole co-operation and encouragement and critical inputs in the preparation of this report. Finally I express my heartfelt thanks to our Lab Instructors, colleagues, friends and my dear Parents for giving me valuable advice and support throughout my project work

REFERENCES

[1].Johannes Schneider 'Detecting Plagiarism based on the Creation Process'
[2].Nouman Azam'Text Summarization using Rough Sets'
[3].Rini.J 'Study on Separable Reversible Data Hiding in Encrypted Images',International Journal of Advancements in Research & technology,Volume 2,Issue 12,December-2013
[4].C.Anuradha,S.lavanya,'Secure and Authenticated Reversible Data Hiding in Encrypted Images',International Journal of Advanced Research in Computer Science and Software Engineering Volume 3,Issue 4,April 2013 ISSN:2277 128X
[5].Lalit Dhande,Priya Khune,Vinod Deore,Dnyaneshwar Gawade,'Hide Inside-Separable Reversible Data Hiding in Encrypted Image',International Journal of Innovative Technology and Exploring Engineering(IJITEE)ISSN:2278-3075,Volume-3,Issue-9,February 2014
[6].Xinpeng zhang,'Separable reversible data hiding in encrypted image', Ieee transactions on information forensics and security, vol 7, no.2,April 2012