TEXT BASED PLAGIARISM DETECTION

Manav Bagai[1], Vibhanshu[2], Siddharth Gupta[3], Rashid Ali[4]
Department of Computer Engineering,
ZakirHussain College of Engineering and Technology, AMU Aligarh.

Abstract: This paper introduces a computer based plagiarism detection technique which combines the functionality of substring matching and keyword similarity to give more accurate results. To make the algorithm more efficient clustering is done by ranking the documents in which LCS (Longest Common Subsequence Algorithm) based algorithm is used.

I. INTRODUCTION

Plagiarism is an unacknowledged act of copying someone's work. Technically, as described by Wikipedia [10], it is "wrongful appropriation" and "stealing and publication" of another author's "language, thoughts, ideas, or expressions" and the representation of them as one's own original work". It is serious problem in academics now day. Some types of plagiarism are:

- Direct Plagiarism: It is word by word copying of someones work without giving any acknowledgement to the document from where the person has copied.
- Self Plagiarism: It is copying of own previous work without giving any reference of it.
- Accidental Plagiarism: It is unintentional copying of similar phrases, words or sentences from a document without giving any reference.

Therefore, plagiarism must be detected to prevent stealing of data. Plagiarism detection can be intrinsic and external. In external plagiarism detection, the reference document is compared with all documents such as documents on web or any other database while in intrinsic plagiarism detection the reference document is compared on with external documents. So, to detect plagiarism one can use either manual or computer based techniques. As manual techniques to detect plagiarism are difficult to implement so there is a need to develop computer based techniques which would efficiently detect plagiarism. Several work has been done towards plagiarism detection techniques. Various algorithms like Substring Matching, Keyword Similarity, Fingerprinting, etc. are currently used for detecting text based plagiarism.These algorithms have their own pros and cons. We have presented an algorithm in this paper which is more efficient than the traditional text based plagiarism detection algorithms. We have used the following techniques to detect plagiarism:

-Clustering of similar documents by using longest common substring algorithm.

-Substring Matching.

-Keyword Similarity.

If all documents present are compared with the reference document it would require a lot of time. So to save time clustering based plagiarism detection technique [1] is used in which a cluster or group of similar kind of documents is created among which document is compared. These similar documents are then compared by other plagiarism detection techniques like substring matching to get more efficient result of plagiarism. We have used Substring Matching and Keyword Similarity method for this purpose. In Substring Matching method for text based plagiarism detection, the documents to be compared are broken into several substrings and are stored in different list. The document is divided using any indicator like '.' , ',' , '?' etc. Then, the lists are compared with each other. It plays an important role in detecting plagiarism in application source codes. In Keyword Similarity method for text based plagiarism detection, a list of keywords are given. The documents to be compared are broken on to substrings again on the basis of keywords and based on that the similarity between the documents is calculated.

II. BACKGROUND

Many attempts have been made in the past to detect plagiarized documents. In this section, we will discuss few of them. Most plagiarism detection techniques utilize string-processing algorithms i.e. these methods are used to find the occurrence of the identical string within the document.

Sudhir et al.[4],have proposed plagiarism detection method in which Temporal Difference learning technique isused. Temporal Difference learning is used to improve the speed of system for retrieving the data from database. Also the system improves accuracy of plagiarism detection. They have firstly separating every statement in the document and then Stanford Parser is used for tree formation of sentences after this all sentences are compared with the local and global database to detect plagiarism.

Tashiro et al. [6] introduce EPCI, which is a tool for finding copyright infringement texts. Given a potential plagiarized document D, EPCI extracts several sequences of words, i.e., seed text, and generatesqueries that retrieve a set of Web documents W thatcould be the source of the content of D. Hereafter,EPCI computes the similarity between D and the documents in W. The higher the similarity value betweenD and any document in W, the more likely that in fringement has occurred.

Khmelev and Teahan [5] use the R-measure to recognize plagiarized documents. The R-measure addsthe lengths of the substrings in a given document thatare included in another document in a collection. Byconsidering the normalized R-measure value, it is possible to establish the "repeatedness" of a documentwith respect to others, which establishes the degree ofplagiarism in the corresponding documents.

Metzler et al. [7] establish several levels of similarity amongdocumentsto identifythose that are exactcopies of a given document D, as well as the ones thatare modified versions of D. In accomplishing the task, Metzler et al. [7] first determine the similarity between sentences within any two documents, and basedon the sentence-to-sentence similarity score, the overall similarity value of the documents is determined.

Leung and Chan [8] propose using a natural language processing method to facilitate the detection ofplagiarized documents, not only among the ones created by "cut and paste", but also documents in whichboth the text and the structure of their original sentences are altered while the content of the documentsare intact.

Maria Soledad Pera et al. [9] proposed similarity based plagiarism detection tool, SimPaD, which relies on pre computed word-correlation factors for determiningsentence-to-sentence similarity values that yield the degree of resemblance of any two documents to detectthe plagiarized one, if it exists. SimPaD is designedto detect (non-)plagiarized text documents, which aredigitalized and posted online, using a collection ofWeb documents which includes the original documents of their plagiarized versions. SimPaD, whichcan handle various plagiarism methods such as sub-stitution, addition, and deletion of words in sentences,as well as sentence splitting and merging, provides the users a visual representation of sentences in a sourcedocument that are paraphrased in its plagiarized version.

III. PROPOSED WORK

Our main aim is to develop an efficient algorithm to determine text based plagiarism. We have developed a plagiarism detection application for detecting external plagiarism, in which clustering based on longest common subsequence (LCS), keyword similarityand substring matching algorithms [2] are used. We have implemented a prototype it by using C++ and Python as a programming language.

A. FLOW OF THE SYSTEM

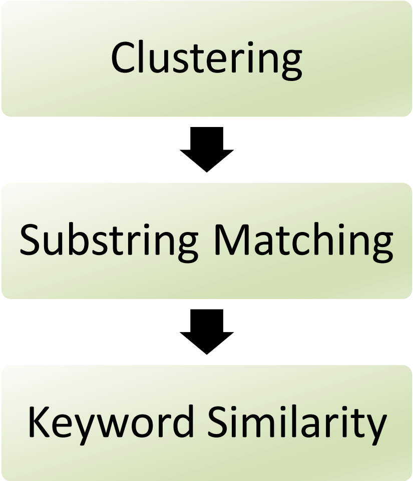


Fig. 1. Flow of the system.

Fig. 1, illustrates the general outline how the whole algorithm works which start from clustering which develop a cluster of similar documents on which substring matching and at last keyword similarity is applied to detect plagiarism. Detailed outline of each process will be given below.

B. CLUSTERING

The **longest common subsequence** (or **LCS**) algorithm[11] finds the longest string between two given strings that are common between the two groups and in the same order in each string. To add to the functionality and accuracy of above algorithm clustering is proposed by us. As there are many documents to be compared so this may take a lot of time. So, to solve this problem clustering is used. Here a cluster is created which mainly contain those files which are similar to the document to be compared by Longest Common Subsequence Method. The steps are shown in Fig. 2. Let the two documents compared be X and Y where X is reference document and Y is document to be compared having length m and n respectively. From longest common subsequence method we will get the length of lowest common subsequence, let it be lcs. To find the similarity between the documents following method [3] is used:

Let the variables R and S be given the following values:

R=lcs/m

S= lcs/n

Then we will find F which is equal to:

$F= (1+B^2)*R*S/(R+B^2S)$ (1)

Where,

B= S/R

The document which is having more value of F, as given in (1) is more similar. If documents are completely same we get the value of F equal to 1. A cluster is made for documents which are more similar to reference document so that next step is applied to only those document.

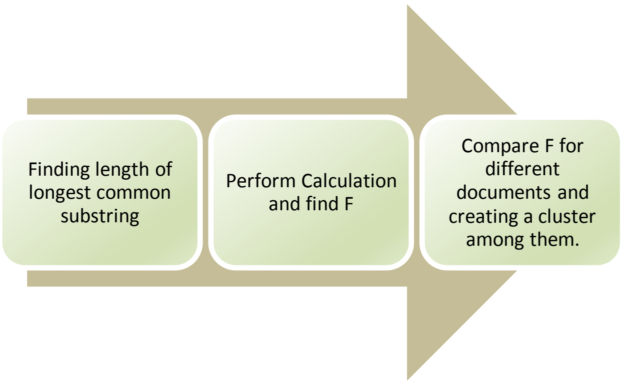

Finding length of longest common substring
Perform Calculation and find F
Compare F for different documents and creating a cluster among them.

Fig. 2. Steps for Clustering

C. SUBSTRING MATCHING AND KEYWORD SIMILARITY

After getting the clusters of the similar documents using LCS, they are no compared using substring matching technique. In this method, we will break the strings from both the documents into substring based on '.', '!', '?'. Then, we will compare all the substrings from both the documents (the reference and the documents from the cluster). If the substrings are found similar, we will increase the plagiarism count. After this, we will apply keyword similarity method, we will ask for a keyword of the document and then using that keyword we will find the sentences from both the documents with that keyword. Fig. 3, illustrates this process.

Then, we will compare those sentences again and if found same we will add them to the plagiarism set. This is shown in the Algorithm [5] below:

```
Terms Used:
Suspected document - Q;
Reference document - D;
Keywords, K – {k1, k2, …kn};
Sentences in Q - {q1,q2,…qn};
Sentences in D- {d1, d2,…dn};
Plagiarism set - P=Null;
Temporary List - One, Two;

Input: Q, K.

//Substring Matching
For Q
Separate sentences Q= {q1,q2,….,qn};
For every q in Q,
Compare with reference document D,
If (q==d)
Add sentence to plagiarized set P,
Update result. P=P+q;
End if
End for
End for

//Keyword Similarity
For every q in Q,
For every k in K,
If(q==k)
Add q to One
         End if
End for
End for
For every d in D
  For every k in K
If(d==k)
Add d to Two
End if
End for
End for
For o in One
For t in Two
If(o==t)
 Add sentence to plag. set P,
Update result. P=P+o;
End if
End for
End for

If (P==NULL)
Display "document is plagiarism free"
Else
Display, set P as a plagiarized sentences.
End else
```

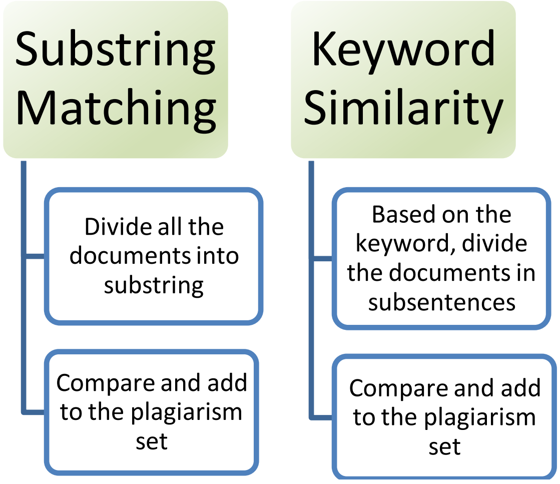

Substring Matching
Divide all the documents into substring
Compare and add to the plagiarism set
Keyword Similarity
Based on the keyword, divide the documents in subsentences
Compare and add to the plagiarism set

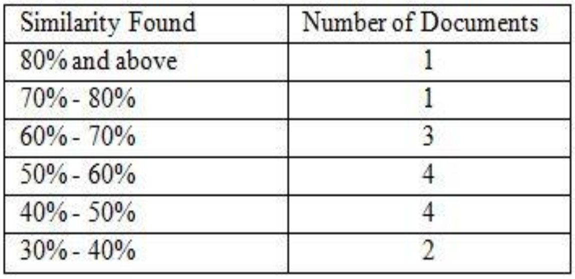
Fig. 3. Substring Matching and Keyword Similarity Process

The process of finding similarity is explained by the example given below: Let a document, say 'file1' is compared with another document say 'file2' selected from the cluster. For example: Let the content of file1 be: "India, officially the Republic of India is a country in South Asia. It is the seventh-largest country by area, the second-most populous country with over 1.2 billion people, and the most populous democracy in the world." Let the content of file 2 is: "China, officially the Republic of China is a country in South Asia. It is the seventh-largest country by area, the second-most populous country with over 1.2 billion people, and the most populous democracy in the world." The similarity found in file1 when compared to file2 is found to be 66.67%.

IV. EXPERIMENTAL RESULTS

In this system we have used clustering based on LCS algorithm to improve performance of the system. The system gives result in terms of total percentage the document isplagiarized. A document is given as an input for which plagiarism is to be checked. Also the documents are to be given as input from which the above document is to be compared. We have performed our experiment by giving 50 documents besides the document in which the plagiarism is to be checked as input, out of which 15 documents having the maximum value of F are selected. Then the substring matching and keyword similarity hybrid algorithm is applied on these 15 documents and similarity between these documents and the above document is found. The similarity of the source document with these 15 documents is shown in Table 1.

Table 1: Similarity found with documents



Similarity Found	Number of Documents
80% and above	1
70% - 80%	1
60% - 70%	3
50% - 60%	4
40% - 50%	4
30% - 40%	2

Here the content of file 1 is:
“Boulder's Human Relations Commission will hold a public hearing on a living wage for employees in the city Thursday evening at the West Senior Center. The hearing is an opportunity to bring community members together to begin a "two-way conversation" about wages and the ability to meet living needs in Boulder, city spokesman Patrick von Keyserling said. The city defines a living wage as "the wage that can meet the basic needs to maintain a safe, decent standard of living within the community," according to a news release. Colorado's minimum wage is currently $8 and will increase to $8.23 on Jan. 1. The federal minimum wage has been $7.25 since 2009. The city has not yet calculated a living wage to present, said Boulder community relations spokeswoman Carmen Atilano. A 2011 study from the University of Washington's Center for Women's Welfare found a single adult would need to make an hourly wage of at least $11.60 to live self-sufficiently in Boulder County. The Colorado Legislature currently limits cities from enacting a minimum wage that is higher than the state's minimum wage, and the hearing will bring up the question of whether the city should consider requesting the state repeal that statute in order to establish a living wage, Atilano said. "It's really in everybody's interest to pay people a fair amount so they're not dependent on public services and can provide for families," said former state Rep. Claire Levy, now executive director of the Colorado Center on Law and Policy. The five person council-appointed Human Relations Commission advises the City Council and wants to recommend that the council consider the issue of a living wage be added to its work plan, Atilano said. This will be the second hearing about a living wage in Boulder this year. The first was conducted Sept. 3 and was hosted by the Human Relations Commission, Boulder Chamber of Commerce, Latino Chamber of Commerce of Boulder County and the League of Women Voters of Boulder County.”
Let the file 2 is:
“Boulder's Human Relations Commission will hold a public hearing on a living wage for employees in the city Thursday evening at the West Senior Center. Colorado's minimum wage is currently $8 and will increase to $8.23 on Jan. 1. A 2011 study from the University of Washington's Center for Women's Welfare found a single adult would need to make an hourly wage of at least $11.60 to live self-sufficiently in Boulder County. The five person council-appointed Human Relations Commission advises the City Council and wants to recommend that the council consider the issue of a living wage be added to its work plan, Atilano said. Boulder police investigating phone scammers pretending to be officers. Boulder police are investigating a phone scam in which callers pretending to be officers ask their targets for money to dismiss arrests warrants or fines, according to a news release.”
Keywords Entered: ‘Boulder’, ‘Human Welfare’.
Plagiarism Detected: 77.777%
Sentences that are found similar: -23 on Jan. Boulder's Human Relations Commission will hold a public hearing on a living wage for employees in the city Thursday evening at the West Senior Center.
-60 to live self-sufficiently in Boulder County.
-The five person council-appointed Human Relations Commission advises the City Council and wants to recommend that the council consider the issue of a living wage be added to its work plan, Atilano said.
-Colorado's minimum wage is currently $8 and will increase to $8.
-A 2011 study from the University of Washington's Center for Women's Welfare found a single adult would need to make an hourly wage of at least $11.

V. CONCLUSION

We have proposed a new plagiarism detection method which is faster and more efficient than the traditional algorithm[4],as it first forms a cluster of the more similar documents to make algorithm faster and then apply a hybrid algorithm of substring matching and keyword similarity on the documents that are there in the cluster to get more accurate results. The algorithm could also be extended further for source code plagiarism detection. The paper does not focus on plagiarism reported in other forms of content e.g., if the original content is represented in textual form and the user has represented in tabular form or an images, which is left for future extensions.

REFERENCES

[1] Du Zou, Wei-jiang Long, Zhang Ling “A Cluster Based Plagiarism Detection Technique” for PAN , CLEF, 2010.

[2] Sangeetha Jamal, “Plagiarism Detection Techniques”, Cochin University of Science and Technology, Cochin 682022, 2010.

[3] Chow Kok Kent, Naomi Salim “Features Based Text Similarity Detection”, Faculty of Computer Science and Informatics System, University Teknologi Malaysia, 81310 Skudai, Johor Malaysia, Journal of Computing, vol 2, issue 1, January 2010.

[4] Sudhir D. Salukhe, S. Z. Gawali, “A Plagiarism Detection Technique Using Reinforcement Learning”, International Journal of Advanced Research in Computer Science and Management Studies, vol. 1,issue 6, November 2013.

[5] D. Khmelev and W. Teahan, A Repetition Based Measure for Verification of Text Collections and for Text Categorization, in: Proceedings of ACM International Conference on Research and Development in Information Retrieval (SIGIR), 2003, pp. 104–110.

[6] T. Tashiro, T. Ueda, T. Hori, Y. Hirate, and H. Yamana, EPCI: Extracting Potentially Copyright Infringement Texts from the Web, in: Proceedings of the World Wide Web (WWW), 2007, pp. 1151–1152.

[7] D. Metzler, Y. Bernstein, W. Croft, A. Moffat, and J. Zobel, Similarity Measures for Tracking Information Flow, in: Proceedings of ACM International Conference on Information and

Knowledge Management (CIKM), 2005, pp. 517–524.

[8] C. Leung and Y. Chan, A Natural Language Processing Approach to Automatic Plagiarism Detection, in: Proceedings of ACM Conference on Information Technology Education (SIG-ITE), 2007, pp. 213–218.

[9] Maria Soledad Pera and Yiu-Kai Ng," Sentence-Based Plagiarism Detection Tool on Web Documents", Web Intelligence and Agent Systems: An International Journal 0, 2009

[10] http://en.wikipedia.org/wiki/Plagiarism

[11] http://en.wikipedia.org/wiki/Longest_common_subsequence_problem