PLAGIARISM DETECTION PARADIGM FOR WEB CONTENT USING SIMILARITY ANALYSIS APPROACH

Mr. M. Ramaraj
Research Scholar,
Department of Computer Science,
NGM College,
Pollachi, Tamil Nadu, India

Dr. Antony Selvadoss Thanamani
Associate Professor and Head,
Department of Computer Science,
NGM College,
Pollachi, Tamil Nadu, India

Abstract— In this paper, we discuss the plagiarism detection paradigm for web content using similarity analysis approach to compare with the two string or word or document. Heterogeneity is the World Wide Web hypertext documents continually growing information sources, unavailability of parameter in the system, the automatic discovery, composition, and web-based information is the most challenging task to manage. In this study, while comparing the text or word disestablishment of vocabulary is encountered. To overcome this issue a method is proposed which incorporates the cosine metric factor to illustrate the relevance among documents while comparing with text or word or group of string. This study reveals efficient detection plagiarism through similarity analysis.

Keywords— Plagiarism Detection, Distance Matrix, Similarity Analysis.

I. INTRODUCTION

Plagiarism is the one of the forms of misuse of academic activities has increased rapidly in the quick and easy to data access and information through electronic documents and the internet. We mean the written we talk about the plagiarism detection for text written by others where they are re-adjust the text to format adding or deleting without any citation or reference.

Some types of plagiarism detection methods such as copy and paste is the most common, redrafting or paraphrasing of the text and plagiarism is the one of the idea for the text processing in the computer science.

Plagiarism is to transform from one language to another and we can used for the many methods that use plagiarism. It is a serious problem in the computer science [2]. In addition, students are becoming more comfortable with cheating. A recent study found that 70% of students admit to some plagiarism, with about half being guilty of a serious cheating offense on a written assignment.

In addition, 40% of students admit to using the "cutand- paste" approach when completing their assignments [1]. The key and main issue in plagiarism detection field is how to differentiate between plagiarized document and non-plagiarized document in effective and efficient way.

These distances are used to change of applications ranging from DNA analysis to detect the theft.

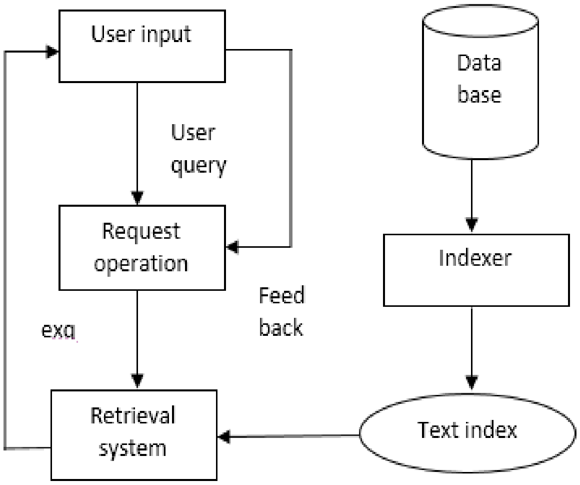

figure 1: general structure for the information architecture

Similarity measures can be comparing with single input sequence to several other representative target sequences. It can be classifying the sequence by finding the most similar or closest target sequence classifies the large number of input. Similarity measures can be computed between several sequences to form a similarity matrix [3]. For example, given k matrix as (k X k) for the cosine factor matrix can be constructed with ijth element contain the similarity analysis Relationship with a precise set of techniques or space and evaluate the pair wised material in position between the pair of object to be closed [12]. Similarity activities planned a solidarity action or cosine similarity of this similarity is to be calculating from the correlation coefficient in a jaccard

Corresponding Author: Mr. M. Ramaraj, NGM College, Pollachi, Tamilnadu, India.

ISSN 2348 – 9928
Doi:01.0401/ijaict.2014.04.01 Published on 05 (1) 2016

similarity [5]. Euclidean distance besides fundamental entropy pair wise set of object to measure the distance and calculate the similarity values or vector values. Given the diversity of similarity to be calculating the distance to detect the effectiveness in the writing manuscript and perform the cluster concept will not clear information.

II. RELATED WORK

2.1 Plagiarism Detection Overview

Theft detection, now order to select the text has a right to discriminate non-plagiarized documents is an important aspect to consider. [2] The vocabulary changes, the level of similarity among books or words separated by a couple of features that can be used in order to find the frequency of theft document.

It is a suspect in the theft of the candidates in order to find the piece that captures the style of the whole document. This approach saves the cost comparison process, but there's no mention of it as a source of potential plagiarized text pieces. In cases where it is considered a reference corpus, based on various aspects of the search process for the different features.

They compare with the relative similarity of the local unit. Copy of the appropriate penalty for insertion into word or string to be process of removal and rewording [6]. Some authors "that can occur in many contexts," unacknowledged copying documents or programs, "theft is defined as: a company's competitive advantage in the field may be obtained; education academics to publish their research in advance, you can request that their colleagues."

Theft detection techniques, there are many programs to view documents before applying for one. Documents such as spacing between words, periods (full stops) with disregard, as an array of characters in the character-based representation, in simplest form, some document between predictors of report and lines [4].

2.2 Plagiarism tools

Plagiarism detection tools have been used for the author in particular tools is use on the student, research scholar and etc.

The top plagiarism tools are available websites address as given below [6][7][8][9].

www.plagiarism.net

www.dupli checker.com

www.ithenticate.com

www.plagiarismchecker.com

III. SIMILARITY MEASURING TECHNIQUES

Plagiarism detection method which has proven to be successful in a number of applications is finding the overlap matching in the string and substring of the length ≥ n, the longer n becomes the number string and same sequence of n is taken [10]. A similarity function is used to capture the degree of overlap between the two strings and used to variety of different similarity measures to the string. The formula as following:

$$similarity(SA, SB) = \frac{\sum_{i \in T} loni \times \log(leni+1)}{|S_A|} \ldots\ldots\ldots\ldots\ldots\ldots..1$$

The formula 1 is represented by similarity between two set of string as S_A and $S_{B.}$ T is a weighted function depended on the length as leni.

Classification of similarity measures as following:

- Cosine similarity measure
- Jacard similarity measure
- Euclidean similarity measure
- Metric similarity measure

3.1 Distance Metric

It is different types of distance metric that can be used to compare the input and target string value. Let d_{ij} = d (x_i, y_j), it represent such value is a distance metric between an input and target value of $x_{i,}$ $y_{j.}$

- ***Cosine similarity measure***

It is represented to the term of vectors is to calculate and find similarity between two string correspond to the correlation between the vectors. This is quantified as the cosine of the angle between vectors, that is, the so-called cosine similarity[9]. It is most popular techniques for cosine matrix and calculate the similarity between the documents.

- ***Jaccard similarity measure***

The sample data sets for the unity, diversity is to compare with statistic to be calculating and to measure the similarity between the limited sample sizes of volumes.

$$j(A,B) = \frac{|A \cap B|}{|A \cup B|} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 2$$

- Euclidean similarity measure

The cosine of the angle between the inner product spaces that measure similarity between two vectors. The cosine of 0’s and 1’s, cosine distance is a team often in the positive space [11]. It is rule that is measurable between two points is a normal place.

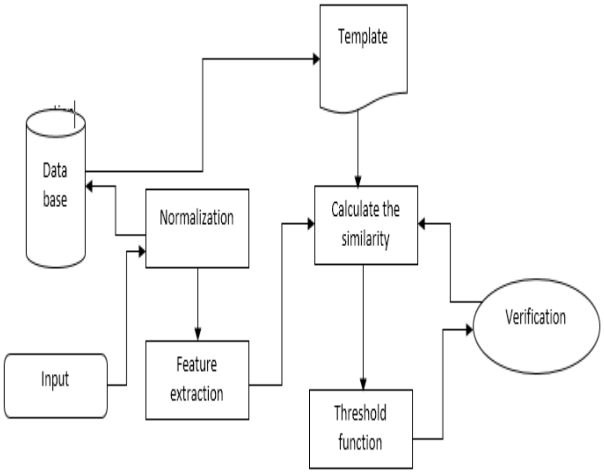

figure 2: the proposed method for the plagiarism detection

- ***Cosine similarity measure***

It is represented to the term of vectors is to calculate and find similarity between two string correspond to the correlation between the vectors. This is quantified as the cosine of the angle between vectors, that is, the so-called cosine similarity [9]. It is most popular techniques for cosine matrix and calculates the similarity between the documents.

- ***Jaccard similarity measure***

The sample data sets for the unity, diversity is to compare with statistic to be calculate and to measure the similarity between the limited sample sizes of volumes.

$$j(A,B) = \frac{|A \cap B|}{|A \cup B|} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 3$$

- ***Euclidean similarity measure***

The cosine of the angle between the inner product spaces that measure similarity between two vectors. The cosine of 0’s and 1’s, cosine distance is a team often in the positive space [11]. It is rule that is measurable between two points is a normal place.

- ***Metric similarity measure***

The distance between any two points must be nonnegative, that is, d(x, y) ≥ 0.

IV. EXPERIMENTAL RESULT

The document collection is used to test our algorithm in a cosine similarity dataset. The data has been number documents and the user queries to display the figure 1 for given the information. It is implemented to the algorithm as for the MATLAB, and find the cosine similarity we have used TMG: as for the Text to Metric generator using MATLAB toolbox.

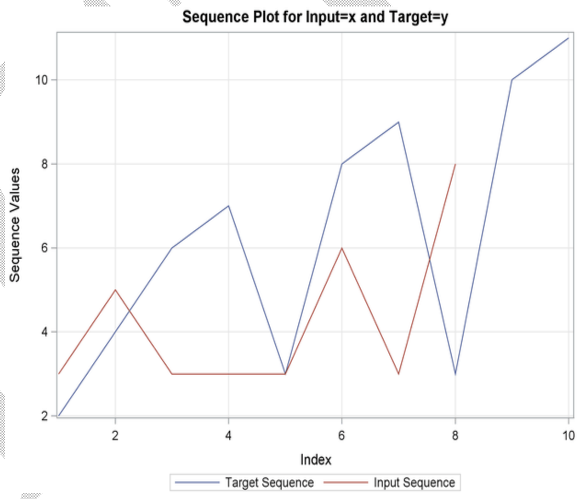

Fig 3: compare with the two values to measure the similarity of this chart.

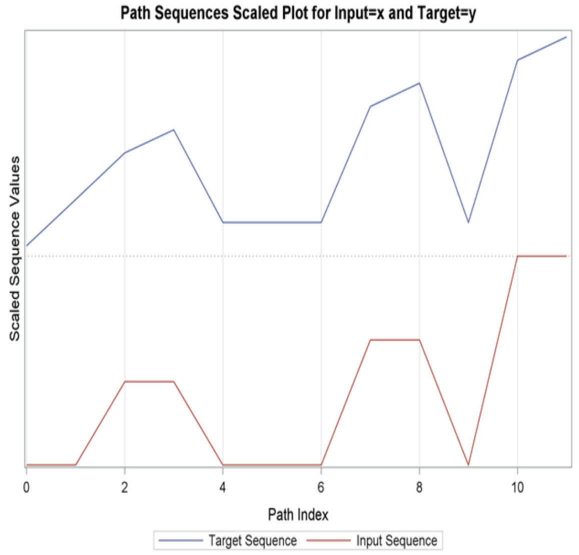

Fig 4: calculate the distance between cosine similarities.

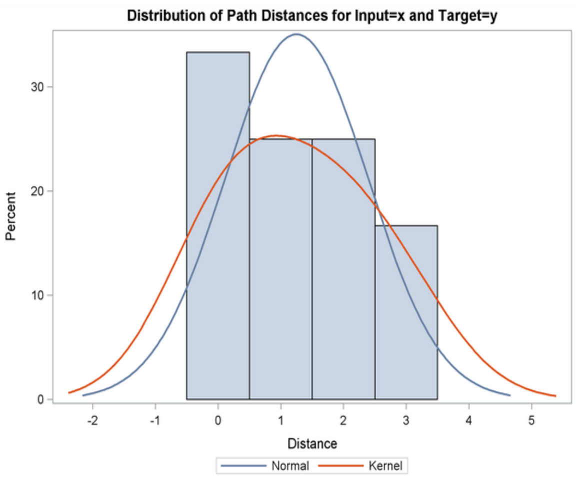

Fig 5: distance and plot

V. CONCLUSION

This paper incorporates the similarity analysis of two string or words or sentence or document file which is to be compared the cosine similarity measures which is helpful for the similarity among the documents. The degree of plagiarism detections has considerably elevated through similarity analysis. Future work could implement pattern matching along with association rule mining to detect plagiarism.

References

[1] P.Kalyan Chakravarthy , J.Bindu Kavya, K.Sireesha, D.Mounika Plagiarism Detection Considering Frequent Senses Using Graph Based Research Document Clustering (IJCSIT) Vol. 5 (1) , 2014, 789-791 ISSN: 0975-9646.

[2] A. Aho, M. Corasick. Efficient String Matching: an Aid to Bibliographic Search. Communications of the ACM, vol. 18(6), 1975, p. 333-340.

[3] B. Baker. Parameterized Duplication in Strings: Algorithms and an Application to Software Maintenance. SIAM Journal onComputing, vol. 26(5), 1997, p. 1343-1362.

[4] S. Burrows, S. M. M. Tahaghoghi, J. Zobel. Efficient Plagiarism Detection for Large Code Repositories. Software Practice & Experience, vol. 37(2), 2007, p. 151-175

[5] Lukashenko, R., Graudina, V. and Grundspenkis, J, "Computer-Based Plagiarism Detection Methods and Tools: An Overview", International Conference on Computer System and Technologies- CompSysTech'07 ACM ISBN: 978-954-964-50-9, 2007.

[6] http://www.turnitin.com (accessed on 15th August 2012).

[7] http://www.duplichecker.com (accessed on 15th August 2012).

[8] http://www.articlechecker.com (accessed on 15th August 2012).

[9] http://plagiarismcheckerx.com (accessed on 15th August 2012).

[10] Jurriaan Hage, Peter Rademaker and Nike van Vugt , "A comparison of plagiarism detection tools", Technical Report UU-CS-2010-015, 2010.

[11] Abeer Al Jarrah, Izzat Alsmadi and Zakariya Za'atreh. "Plagiarism Detection based on studying correlation between Author, Title, and Content", InternationalSConference on Information Communication System (CICS), May 22-24, 2011.

[12] Ahmed Hamza Osman, Naomie Salim, Mohammed Salem Binwahlan Plagiarism Detection Using Graph-Based Representation Journal of Computing, Volume 2, Issue 4, April 2010, ISSN 2151-9617.