ISSN No. 0976-5697

Volume 5, No. 3, March-April 2014

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Effective Analysis of Nearest Duplicate Text Document using Fuzzy Clustering Method

Nancy Jasmine Goldena
Research Scholar,
Department of Computer Science,
Mother Teresa Women's University,
Kodaikanal- 624101, India.

Dr. S.P. Victor
Associate Professor and Head & Director of the Research Center, Department of Computer Science,
St.Xavier's College,
Palayamkottai-627 002, India

Abstract— In this research article, the descriptive and validating of the most popular fuzzy clustering methods for the detection of near-duplicate text documents are framed. The fuzzy based cluster algorithm analyze the resemblance of various information that available in online through the applied two-stage protocols. In first stage, the fuzzy clustering algorithm analyze the duplicate content of the image through RGB that debased with Euclidean distance metric and in the second stage, the similarity syntax of the text mining are verified through the parametric view based on the time series for sample for each document. It is proposed to develop methods of assessing the adequacy of the criteria changes in ontologies for the development of fuzzy state space.It involves the analysis and identification of web duplicate resources through Fuzzy based method. The theoretical results are confirmed recommendations for practical use. The algorithm for filtering near-duplicate documents discussed here has been successfully implemented and it shows better accuracy than other exisiting alogirthms.

Keywords—: Fuzzy sets, Soft computing, Web mining, Information retrieval, Duplicate Document.

I. INTRODUCTION

The use of computing and communication - ventilating systems for more and more into the activities of all modern companies. Virtually every company now has not only its own website, and a web-application, adapt all mobile phones and smartphones. The function of most companies completely depends on web-application possibilities are continuously very fast growing. Thus, the problems associated with insufficient accurate performance, and there will be impact penetrate , leading to the failure of customers use certain applications . In this context , the relevance method adopted the qualitative load testing, which should be mandatory to ensure the stability of the application. The problem of near-duplicate detection is one of the most important and difficult tasks of web data analysis and retrieval of information on the Internet. The urgency of this problem is determined by a variety of applications in which it is necessary to consider the "similarity", for example, text documents, that improve the quality of the index and search the archives by removing redundant information, and association news stories in stories based on the similarity of these messages for content and spam filtering (both mail and search), and the establishment of copyright infringement in the illicit copying of information (the problem of plagiarism or copyright), and several others in [1].

The main obstacle to the successful solution of this problem is the huge amount of data stored in the databases of modern search engines. This volume makes it almost impossible (in a reasonable time) of its "direct" solution by pairwise comparison of texts in [2]. Therefore, recently a lot of attention is given to developing methods to reduce the computational complexity of algorithms created by selecting different heuristics (eg , hashing certain fixed set of " significant" or words.The aim of this work is the consideration of load testing and analysis of the results of testing conconcrete site using methods of fuzzy logic and comparing fuzzy ontologies and identify their relevance in the process of development that explained in [3].

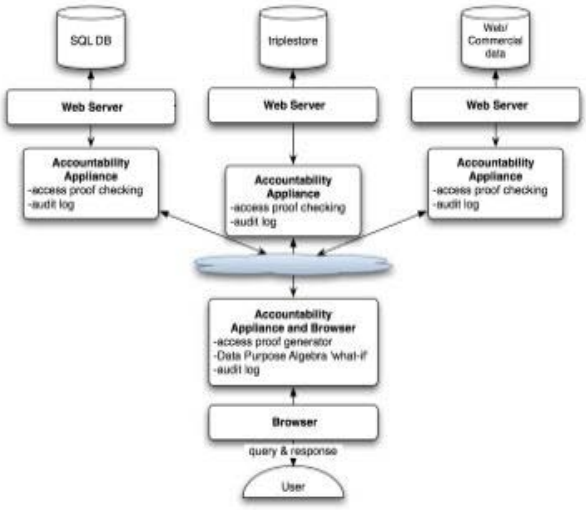

Fig 1.1 Web Access Accountability

This work focuses on proposals document sample set of substrings of the text , the use of fingerprint, etc.). In applying the approximate approaches , a decrease (sometimes very large) index completeness detect duplicates. An important factor affecting the accuracy and completeness of duplicates in the web search problems is the allocation of the content of the web pages with the help of a reliable identification of elements of registration documents and their subsequent removal . In this paper, these issues are not addressed. Finally, another key requirement for quality detection algorithms fuzzy duplicates is their resistance to a "small" changes in the source documents and the ability to confidently handle short documents

II. LITERATURE SURVEY

According to Zadeh [1], fuzzy logic may serve as the backbone of the Semantic Web, an extension ofthe current Web in which information is given well-de8ned meaning, thereby better enabling computers and people to work in cooperation. One of the first studies in the field are finding near-duplicate of Manber [2], and N. Heintze [3]. In these studies, the sample used to construct the sequence of adjacent letters . Fingerprint File ([2]) or document [3] includes all text substrings of fixed length. The numerical value is calculated using the fingerprint algorithm of random polynomials Karp -Rabin [4]. As a measure of similarity between two documents used ratio of the number of common substrings to the size of a file or document. Proposes a number of methods to reduce the computational complexity of the algorithm. U. Manber used this approach for finding similar files (utility sif), and N. Heintze - for the detection of near-duplicate documents (system Koala).

In 1997 A. Broder et al. [5] proposed a new " syntax " method of assessing the similarity between documents based on the representation of the document as a set of all possible sequences of a fixed length k, consisting of adjacent words in [6]. Such sequences have been called " singles ." Two documents are considered similar if their sets of singles significantly overlap. Since the number of singles is approximately equal to the length of the words in the document , i.e. is large enough , the authors have proposed two methods for sampling representative subsets. The first method is left only those singles , whose " fingerprint " , calculated by Karp -Rabin algorithm, without the rest shared by some number m. The main drawback - the dependence of the sample length of the document and, therefore, small size documents (words) to be made or a very short samples or do not have any. The second method selects only a fixed number s singles with the lowest values or fingerprint left all singles , if the total number does not exceed s. Calculated for each chain 84 fingerprint algorithm of Karp-Rabin using a one-to-one and independent functions using random sets («min-wise independent») simple polynomials in [7]. As a result, each document represents 84 singles, minimizing the value of the corresponding function in [8]. The main drawback - the instability to small changes in the content of the document.To overcome this drawback, the original algorithm was modified by the authors, and it was introduced the possibility of multiple random shuffle the main dictionary. The essence of the new improvements consists in the following.

III. METHODOLOGY

The basic idea of the study was that , using different sets of text documents (public collections , e-mail messages , individual Web pages , etc.) , to assess the quality of the most well known , diverse and effective computational standpoint duplicate detection algorithms. Another objective was to develop new methods in this area, taking into account the shortcomings and limitations of existing approaches in [9]. The main indicators of the quality of the algorithms were chosen completeness, accuracy and F- measure. Supposed to compare algorithms on these parameters, as well as to determine their cross-correlation and joint covering different combinations of algorithms for initial set of pairs of fuzzy duplicates in [10].

Performance of each algorithm (pair of near-duplicate) may be aggregated so that for any pair of source documents (docX, docY) specifies a list of algorithms (A1, A2, ...), in terms of which these documents were fuzzy duplicates in [11].. All couples candidates received at least one algorithm can be verified by direct comparison of the text , which will evaluate the accuracy of each algorithm. For a comparative evaluation of the completeness of algorithms can use the idea of a common pool as a maximal set of duplicates found and verified.

IV. FUZZY CLUSTER WEB DOCUMENT CLASSIFICATION

Accurately capturing the overlay topology of a large scale of cluster image classification of spectral information involves soft computing method to classify the objects in the tumor tissue cluster images. The soft computing is a flexible methodology that exploits the fault tolerance for uncertainty through neural networks and fuzzy c-Means Algorithm based on the feature resources available in the spectrum of images. The colors of various combinations that present in the given cluster images are Blue (B), Green (G) and red (R). In spectrum measurement, these three color variation extract the specific spectrum information from the object through the Euclidean distance metric and through the HIS-Model hue (H) describes the pure color in terms of the dominant wavelength and it's given by

$$H = \cos^{-1}\left\{ \frac{\frac{1}{2}[(\mathrm{R}-\mathrm{G})+(R-B)]}{[(\mathrm{R}-\mathrm{G})^2+(R-B)(G-B)]^{1/2}} \right\} \quad (1)$$

Again, the saturation (S) gives amplification of white color debasing the real color of the image and it given by

$$S = 1 - \frac{3*\min(\mathrm{R},\mathrm{G},\mathrm{B})}{(R+G+B)} \quad (2)$$

Then, Intensity measure is the average of all combinations of different object color by [14]

$$I = \frac{1}{3}(R+G+B) \quad (3)$$

In the next stage, the fuzzy C-Means algorithm applied for finding the intra cluster distance though minimizing the objective function which relevant data point for a set of prototypes:

$$J_{FCM} = \frac{1}{2}\sum_{X=1}^{N}\sum_{i=1}^{c} \mu^{m}_{x,i} d^2(z_x, v_i) \quad (1)$$

Here, $\mu_{x,i}$ (x=1, 2,…., N, i=1, 2…, c) is membership value, it denotes fuzzy membership of data point x belonging to class I, V_i (i=1, 2,…, c) is centroid of each cluster and Z_x (x=1,2,….,N) is data set (pixel values in image), m is fuzzification parameter $d^2(Z_x, V_I)$ is Euclidean distance between Zx and Vi, N is the number of data points, C is the number of clusters.

Fuzzy partition is carried on an iterative optimization of the equation (1) based on [12]:

a) Choose primary centroids V_i (prototypes).

b) Computes the degree of membership of all data set in all the clusters:

$$\mu_{x,i} = \frac{\left(\frac{1}{d^2(z_x, v_i)}\right)^{(1/m-1)}}{\sum_{i=1}^{c}\left(\frac{1}{d^2(z_x, v_i)}\right)^{(1/m-1)}} \quad (2)$$

c) Compute New centroids V^1_i:

$$V_i^t = \frac{\sum_{x=1}^{N} \mu_{x,i}^m z_x}{\sum_{x=1}^{N} \mu_{x,i}^m} \quad (3)$$

and update the degree of membership $\mu_{x,i}$ according to the equation.

d) If $\mu_{x,i} <$ □ stop, otherwise go to step 3 (4)

Where □ is a termination criterion between 0 and 1, $d^2(z_x, v_i)$

$$d^2(z_x, v_i) = \| z_x - v_i \|^2 \quad (5)$$

The above ratio provides better experimental results and preserves intelligibility of image together in the cluster image through the Fuzzy C-Means method and analyses with less time variance factor for image classification analysis of different types of duplicate files in [13]. The main problem that the authors set themselves the development of new algorithms for detecting near-duplicate - a significant (2-4 fold) increase in the "fullness" in comparison with existing algorithms, while maintaining the highest possible measure "accuracy". In this method, we experimented with complete sets of singles and methods.

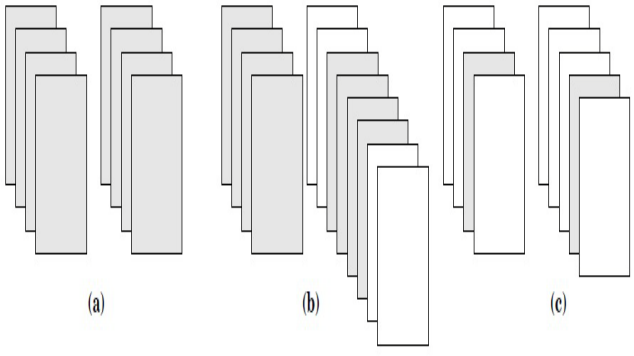

Figure 4.1: Definitions of duplicate documents: (a) the two documents are identical, (b) one document is included in the other, (c) part of one document is included in the other.

The algorithm is as follows :

(a). For each document not calculate recurring singles (can not all but selected) and save the file in a format <single, doc_id, doc_len> in [14].

(b). To construct a chain of identical singles format <doc_id1, doc_len1> <doc_id2, doc_len2> ..., ordered by ascending doc_len. and splits the string into smaller , if the ratio of the lengths of the adjacent greater lengths to less exceeds a certain threshold , defined minimum level of similarity for duplicates (for example, the level of similarity of 0.85 can be virtually no loss of completeness using a threshold 1.15).

(c). Remove duplicates chains and chains that are entirely included in the other . As a result, the number of chains is reduced by hundreds of times , and the remaining chains are overwhelmingly short enough (2 - 10 items) that based on [15].

(d). Documents within the chain compare pairs (for example, by using Perl Similarity function or even easier , with the help of numerical characteristics kakih-nibud additional documents) in [16]. And the comparison is again not the entire chain , but only within a small local neighborhood , defined by the same threshold ratio of the lengths . Therefore, the total number of actual comparisons little.

(e). To avoid duplicate checks in different chains lists have already been processed in the hash pairs save.

V. EXPERIMENT RESULT

Experiments have shown effectiveness proposed in completely automatic method in formation documents and categorization that is based in the applied clustering algorithm for full texts . Software implementation of the method designed for digital libraries as an element of their search engines. Such an element capable of be as independent search mechanism and serve as a means to improve the quality of other search engines , for example, search by keyword.

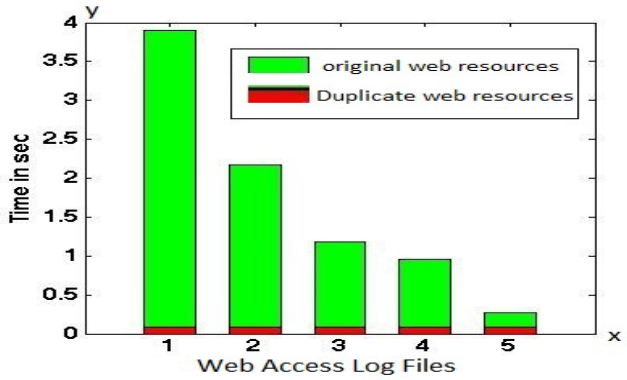

Figure 5.1 Identification of web duplicate resources through Fuzzy based method

It also proposed method can be used in the preparation tools for analyzing the dynamics of scientific technical expertise in electronic document collections. The first attempt to create an information system based on the Collection of documents as a group. The program interface contained flowcharts groups , allowing for the conversion images to other formats (JPEG / PNG) for reducing the size of a partial loss color information , and the resulting scheme can not be used to navigate the same means that the local system.Search for images intended for searching for images similar to the fragment . The input is investigated image , and the output should appear images from the database , the most similar to original. The implementation stage involves careful planning, investigation of the existing system and it's constraints on implementation, designing of methods to achieve changeover and evaluation of changeover methods. Implementation is the process of converting a new system design into operation.

Although, as previously stated, not all can be calculated singles, but only some of them, according to a heuristic Similarity and replaced by simpler means. Algorithm showed almost 100% completeness The duplication content in the document are detected based on the template of syntax. It includes the page details, text font, text captions, text inside the table. When the search of the particular document is finished, the algorithm will not compare the same data again and again to test for inclusion. it improves the efficiency of an algoirthm through fuzzy based methods. Duplicates were constructed from the query by changing the line breaks and/or appending roughly equal amounts of unrelated text to the beginning and end of the document.

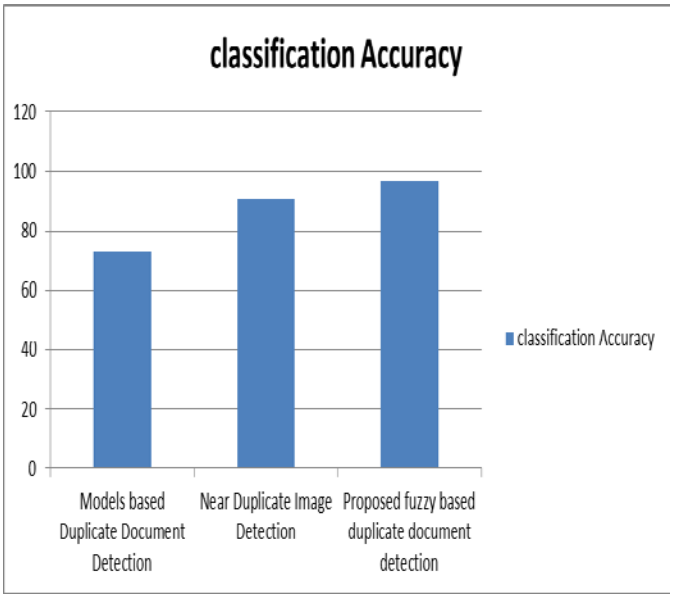

Figure 5.2 Comparative Graph of the proposed system Accuracy in classification based on SAS implementation in [20].

In this method, the obtained experimental results are evaluated through evaluation metrics namely, sensitivity, specificity and accuracy. In order to find these metrics, we first compute some of the terms like, True positive, True negative, False negative and False positive based on the definitions given in Table 5.1. The obtained results depict that the proposed duplicate content detection approach produces better results than the exiting classifier in terms of sensitivity, specificity and accuracy.

VI. CONCLUSION

In this paper, we projected a general method of the algorithms that based on the fuzzy clustering and repeative iteration scheme over the analysis of nearest duplicate text document that are untapped opportunities to increase the completeness of which can be integrated into a single scheme, and try to overcome the disadvantages of each algorithm in this regard. The fuzzy based cluster algorithm analyze the resemblance of various information in the document that contains both images and text that available in online through the applied protocols. In first stage, the fuzzy clustering algorithm analyze the duplicate content of the image through RGB that debased with Euclidean distance metric and in the second stage, the similarity syntax of the text mining are verified through the parametric view based on the time series for sample for each document. Later, the similarity of the text mining are verified throguh the parametric view based on the time series for sample for each document and compared the results with the various existing methods. The comparative results shows the proposed algorithm is efficient to analyze the duplicate document in less time with reduced misclassification than other methods.

In future, this fuzzy clustering algorithm will be implemented and analyzed on text based images to find the similarity.

VII. REFERENCE

[1]. Zadeh, L. A. (1996). Fuzzy logic= computing with words. Fuzzy Systems, IEEE Transactions on, 4(2), 103-111.

[2]. Manber, U. (1994, January). Finding Similar Files in a Large File System. In Usenix Winter (pp. 1-10)

[3]. Heintze, N. (1996, November). Scalable document fingerprinting. In 1996 USENIX workshop on electronic commerce (Vol. 3, No. 1).

[4]. Broder, A. Z. (1997, June). On the resemblance and containment of documents. In Compression and Complexity of Sequences 1997. Proceedings (pp. 21-29). IEEE.

[5]. H. Frigui and R. Krishnapuram, "A Robust Clustering Algorithm Based on Competitive Agglomeration and Soft Rejection of Outliers," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, June 1996, pp. 550-555.

[6]. S. Medasani and R. Krishnapuram, "Determination of the Number of Components in Gaussian Mixtures Based on Agglomerative Clustering" Proceedings of the IEEE International Joint Conference on Neural Networks, Houston, June 1997, pp. 1412-1417.

[7]. S. Medasani and R. Krishnapuram, "Determination of the Number of Components in Gaussian Mixtures Based on Agglomerative Clustering" Proceedings of the IEEE International Joint Conference on Neural Networks, Houston, June 1997, pp. 1412-1417.

[8]. E. Cox, The Fuzzy Systems Handbook: A Practitioner's Guide to Building, Using, and Maintaining Fuzzy Systems, AP Professional, 1994.

[9]. M. Ma,A. Kandel, and M. Friedman, "A New Approach for Defuzzification," Fuzzy Sets and Systems, vol. 111, no. 3, May 2000, pp. 351–356.

[10]. J. Zhang and J.W. Modestino, "A model-fitting approach to cluster validation with application to stochastic model-based image segmentation," IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 12, No. 10, October 1990, pp. 1009-1017.

[11]. J.-M. Jolion, P. Meer and S. Bataouche, "Robust clustering with applications in computer vision", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 13, no. 8, August 1991, pp. 791-801.

[12]. C. V. Stuart, 'MINPRAN: A new robust estimator for computer vision," IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 17, no. 10, October 1995, pp. 925-938.

[13]. M. P. Windham, "Numerical Classification of Proximity Data with Assignment Measures," Journal of Classification, Vol. 2, pp. 157-172, 1985.

[14]. R. J. Hathaway and J. C. Bezdek, "NERF c-Means : Non-Euclidean Relational Fuzzy Clustering," Pattern Recognition, Vol. 27, No. 3, pp. 429-437, 1994.

[15]. P. H. A. Sneath and R. R. Sokal, Numerical Taxonomy - The Principles and Practice of Numerical Classification, W. H. Freeman and Co., San Francisco, 1973.

[16]. Anupam Joshi, S. Weerawarana, and E. Houstis, "On Disconnected Browsing of Distributed Information", Proc. Seventh IEEE Intl. Workshop in Research Issues in Data Engineering, 1997, pp. 101-108.

[17]. Spitz, A. L. (1997, March). Duplicate document detection. In Proc. SPIE, Document Recognition IV (Vol. 3027, pp. 88-94).

[18]. Lopresti, D. P. (1999, September). Models and algorithms for duplicate document detection. In Document Analysis and Recognition, 1999. ICDAR'99. Proceedings of the Fifth International Conference on (pp. 297-300). IEEE.

[19]. Chum, O., Philbin, J., & Zisserman, A. (2008, September). Near Duplicate Image Detection: min-Hash and tf-idf Weighting. In BMVC (Vol. 810, pp. 812-815).

[20]. Wen Zhu, Nancy Zeng, Ning Wang, "Sensitivity, Specificity, Accuracy, Associated Confidence Interval and ROC Analysis with Practical SAS Implementations", Proceedings of the SAS Conference, Baltimore, Maryland, pages: 9, 2010.