Volume 4, Issue 8, August 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Entropy-based Fuzzy Clustering Approach with Radon Transformation on Duplicate Data Detection and Identification

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

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

Abstract— The application fuzzy clustering along with the radon transformation for the detection of near-duplicate text documents are analyzed and developed. By providing a quantitative assessment of indicators of completeness, accuracy and F-measure are analyzed. In the duplicate document detection, many factors are measured to be analyzed that contrasts from the document originality in the search engine. In this method, the syntax of characters and substrings of various sizes are bound under the fuzzy clustering. Then, it evaluated through word correlation computing in a four level approach. Then, the detected article undergoes the transformation to identify the originality of the data in other existing regions that are considered as a crawler. The fingerprint is marked for every unique substring in the document. Later, the third stage involves the identification of similarity through the hash function values. Actually, the hash function is constructed using efficient linear local features. In this article, we introduce the possibility of using fuzzy-hashing to produce fingerprints of files (or documents, etc.). Finally, the processed documents are validated with an F - measure that the computation between various documents through the Recall values.

Keywords— Fuzzy sets, soft computing, Web mining, near duplicate detection & Data cleaning

I. INTRODUCTION

The problem of near-duplicate detection is one of the most important and difficult tasks of web data analysis and information retrieval on the Internet. The urgency of this problem is determined by a variety of applications. It is necessary to consider the "similarity" of the methods. Generally, the text documents improve the quality of the index and search the archives by removing redundant information and it also news reports in association plots on the basis of similarity in content of these messages, 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. The modification of documents online can be very different, for example: creation of mirror sites, document conversion in another format, or editing. Separate item is intentional distortions of the text used by spammers in mass mailings. Direct solution by pairwise comparison of the documents in terms of giant levels of data on the Internet is not possible. There are different methods that reduce the computational complexity for by selecting different heuristics such as hashing fixed set of meaningful words, markov chains of text elements that depleting transformations fingerprint, etc. Then, the resulting hashes (fingerprint documents) are compared and documents are considered similar unless share matched prints are more than a threshold. The paper is structured as follows. The first a brief definition of the duplicates and hash functions and coordinate the pattern lying in the based detection algorithm. The second part is devoted to the proper development of an algorithm in finding duplicates to coordinate pattern. In the third part, the brief theoretical information about effective local linear characteristics (LLP) are reverted. The fourth section is dedicated to developing best hash function on the OS basis LLP. The fifth section presents the experimental detailed studies of the proposed hash function and identifies the optimal parameters of the algorithm. Disarmament duplicate template. The process of spam often degrades the quality of search results and increases the load on the Trouble system. He was recognized as one of the main threats to the modern search engines. [2] By some estimates up to 20 % of all Internet content is search spam [3], the level of search spam extradition of leading search engines is 3-6 % [4]. Search engines use a variety of information to rank pages: Filter page and the site on which it is located; links between pages and sites, etc. There are several types of search engine spam, aimed at discrediting the various algorithms used in search engines. For example, link spam is aimed at cheating reference ranking algorithms, such as Page Rank [5]. This paper investigates methods to counter the other varieties of search spam mass generated by unnaturally texts. In the generation of texts target of spammers is to hit the extradition request with a small amount of relevant pages. To maximize the number of clicks users such requests spammers have to create thousands of pages, each of which must appear on one or more low-frequency queries. This spam is particularly dangerous for the search engines since such pages are likely to fall into the issue.

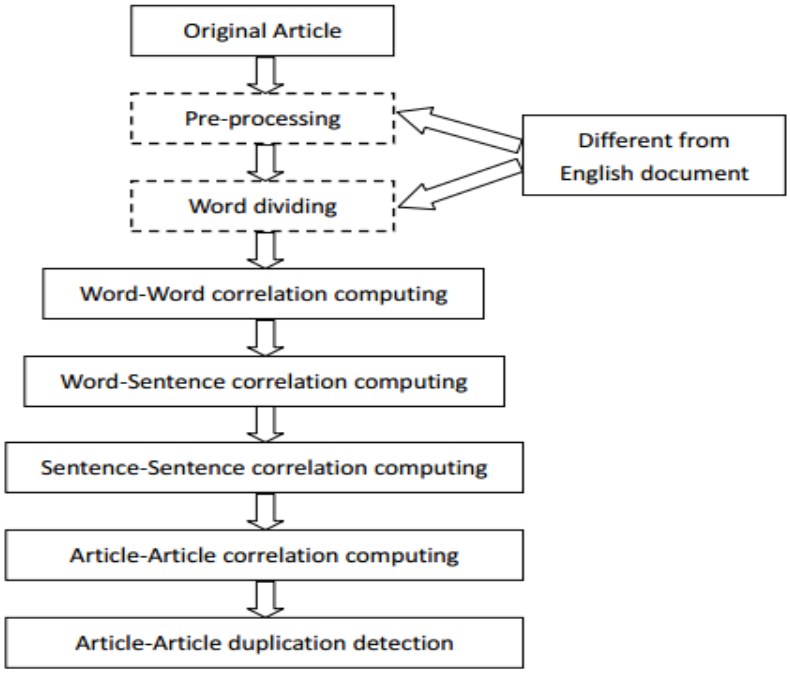

Fig 1.1 The basis of duplicate detection

Spam aimed at discrediting the different algorithm rhythms search engine, and is divided into several areas. Link spam is created to affect weight reference ranking for a particular page. Examples of link spam are special network pages related links. Such structures are aimed at algorithms dairy ranking similar to PageRank [5]. There is a vast class of algorithms, aimed to combat link spam, for example [7].

Another important aspect in the fight against search spam is duplicate detection texts. Overview of methods for the detection of duplicates is given in [6]. The basis of most of the methods duplicates detection is effective fragments copied text on the basis algorithms. This approach is based on an analysis of the statistical characteristics of the texts and the use of machinery, learning to build an automated classifier search spam. Development of this approach analyzed in [9]. In this paper, we use the method of allocation concealment problem for the definition of spam texts

II. METHODOLOGY

Initially, the local marginal information of the data, it is difficult to obtain a perfect result when there's a fuzzy or discrete boundary in the region, and the leading problem is inescapable appeared; Secondly, solving the partial differential equation of the level set function requires numerical processing at each point of the data domain which is a time consuming process; includes all text substrings of fixed length. The numerical value is calculated using the fingerprint algorithm of random polynomials [4]. As a measure of similarity between two documents used the ratio of the number of common substrings to the size of the file or document. Proposes a number of methods aimed at reducing computational complexity of the algorithm. Finally, if the initial evolution contour is given at will, the iteration time would increase greatly, too large or too small contour will cause the convergence of evolution curve to the contour of object incorrectly. Therefore, some modification has been proposed to improve the speed function of curve evolution [10-12]. In the paper, based on the new variant level set method, the edge indicator function was weighted to improve the ability of detecting fuzzy boundaries of the object. In fuzzy logic, fuzzy sets define the linguistic notions and membership functions define the truth-value of such linguistic expressions. Membership function defines the membership degree to a fuzzy set. The basic idea of this approach is fingerprint calculation for I-Match presentation of the content of documents. With this first goal for the initial collection of documents constructed dictionary L, which includes words with average values of the IDF, because such words generally provide more accurate results in near-duplicate detection. Words large and small values of the IDF is discarded.

A. Fuzzy C-Means Clustering

The standard fuzzy, c-means objective function for partitioning cluster is given by

$$J(U,V)=\sum_{i=1}^{c}\sum_{k=1}^{N}u_{ik}^{p}\left\|x_k-v_i\right\|^2 \tag{1}$$

$$\sum_{i=1}^{c}u_{ik}=1\,|\,0\le u_{ik}\le 1,\forall k=1,2,3....N \tag{2}$$

$$0\le\sum_{k=1}^{N}u_{ik}\le N \tag{3}$$

$$F_m = \sum_{i=1}^{c}\sum_{k=1}^{N} u_{ik}^{p} \left\| x_k - v_i \right\|^2 + \lambda(1 - \sum_{i=1}^{c} u_{ik}) \tag{4}$$

In other words, the crawler gets all, the only documents, but The constrained optimization could be solved using one Lagrange multiplier does not stop, it keeps getting URLs that bring new and different from those already explored, leading to pages or documents that are duplicates of others already visited. This means that, quite closely, the path may be suspended long before having explored the entire collected addresses, in the confidence that, after a certain point, few if any pages with new content will be covered.

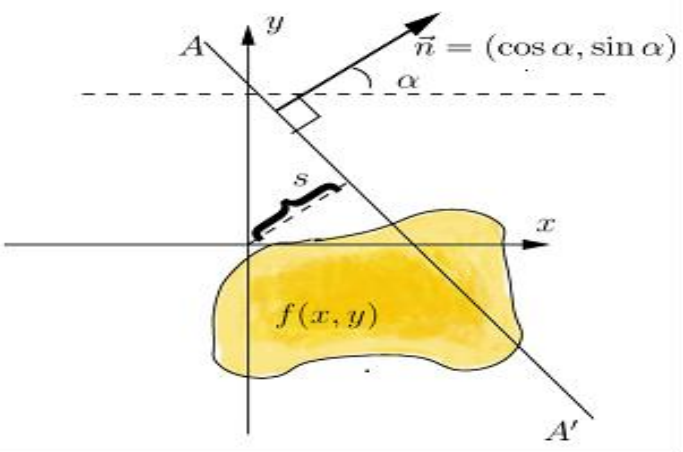

Fig 2.1 Radon Transformation.

$$\frac{\partial F_m}{\partial u_{ik}} = p u_{ik}^{p-1} \left\| x_k - v_i \right\|^2 - \lambda \tag{5}$$

$$u_{ik} = \left(\frac{\lambda}{m} \right)^{\frac{1}{m-1}} \frac{1}{\left\| x_k - v_i \right\|^{\frac{2}{m-1}}} \tag{6}$$

The identity constraint $\sum_{j=1}^{c} u_{jk} = 1 \quad \forall k$ was taken into account,

$$\left(\frac{\lambda}{m} \right)^{\frac{1}{m-1}} \sum_{j=1}^{c} \frac{1}{\left\| x_k - v_j \right\|^{\frac{2}{m-1}}} = 1 \tag{7}$$

This allows us to determine the Lagrange multiplier λ

$$\left(\frac{\lambda}{m} \right)^{\frac{1}{m-1}} = \frac{1}{\sum_{j=1}^{c} \frac{1}{\left\| x_k - v_j \right\|^{\frac{2}{m-1}}}} \tag{8}$$

The zero-gradient condition for the membership estimator can be rewritten as

$$u_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{\left\| x_k - v_i \right\|}{\left\| x_k - v_j \right\|} \right)^{\frac{2}{p-1}}} \tag{9}$$

Where N is the adaptive variable parameter.
For practical studies were selected methods and algorithms for determining fuzzy Duplicate

$$\nabla_{v_i} J = 0$$

The detailed solution depends on the distance function. In the case of the distance, this leads to the expression:

$$2\sum_{k=1}^{N} u_{ik}^{p} (x_k - v_i) = 0 \tag{10}$$

So the following could be immediately obtained

$$v_i = \frac{\sum_{k=1}^{N} u_{ik}^{p} x_k}{\sum_{k=1}^{N} u_{ik}^{p}} \qquad (11)$$

Then, once you have the fingerprint of all the documents, in order to determine if two documents are near duplicates of each other, we can simplify compare the fingerprints, and not worry about the full original document.

B. Fuzzy Clustering Methods

In [1], the author proposed a method construction document signature under construction based on a particular set of statistical document settings, selected on the basis of stability to certain considerations forms of document modifications. E.g. Number of points, commas and capital letter text allows you to fix some - amount of proposals in the text; total length text in characters excluding spaces and stop words gives an overall assessment of the amount; rating the amount and ratio of different frequency words can serve as a kind of substitute for the semantic data analysis etc. This set of data was compiled on the basis of function. "Relevant" assumed to the tallest pair of documents, consisting of no less than 20 words, and the similarity rate, the calculated function Perl String Similarity, is not less than 85% after removal of the tag.

C. The Similar Duplicate Data Detection based on Fuzzy Clustering

In [10] an approach based on the analysis of the compatibility of word pairs to detect unnatural texts are compared. The approach is the assumption that more unnatural texts likely to contain a rare pair of words. In this paper, we propose an algorithm to calculate the proportion of pairs of rare words and show that this feature improves the quality of the definition of search engine spam. In [11], an approach to the definition of unnatural texts, which is based on the hypothesis that unnatural texts cannot simultaneously satisfy all constraints attached natural texts. When learning algorithm, a large number of statistics at signs associated with reading, unity of style and genre features that subsequently combined into an automatic classifier. The approach proposed in this paper is based on the work [11], but significantly expands its based on the account of the properties of the model thematic structure of texts to determine unnatural texts.

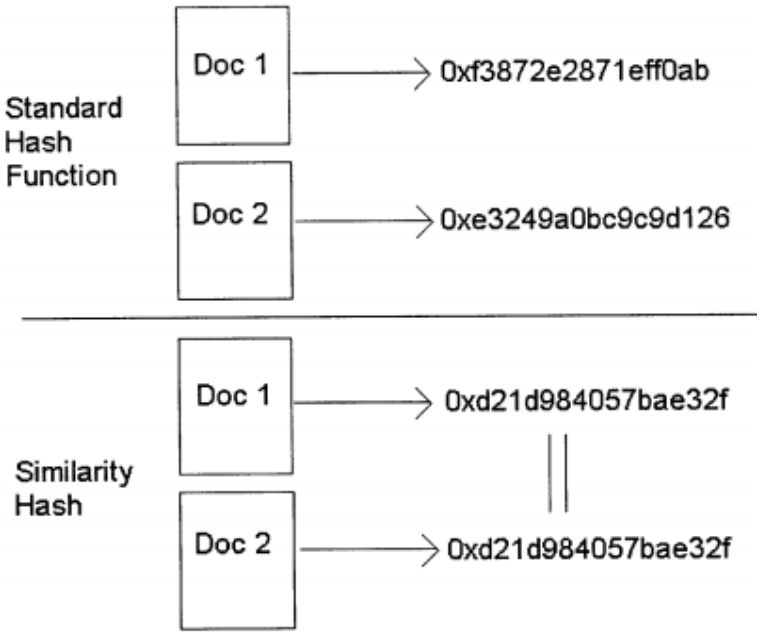

Fig 2.2 Similarity between Two Documents

There are several systems that estimate the resemblance or similarity between two documents. Several of them are derived from the calculation of similarity between vectors, apply where documents are represented by vectors (15). But they are not used in this case; were designed to estimate semantic similarity; vector own on which it rests to use independently of the terms, regardless of position relative to each other. And in any case, your application has some requirements ensembles processing in this context (18). We could calculate the time complexity of the naïve FSA (fuzzy set approach) as follows: Consider 2 articles A and B, both with N sentences and each sentence contains M words. For each sentence we need to check the word-word correlation, if we assume the time consumption for this procedure lasts S seconds. If there are total P articles that match the query term, the entire process may take as long as:

$$T(A, B) = C_P^2 C_M^2 C_N^2 S$$

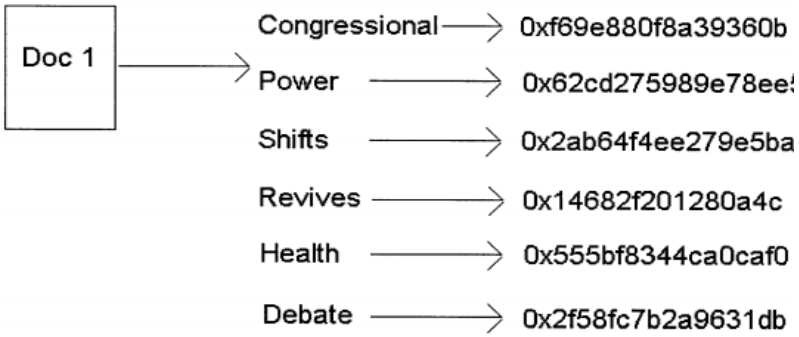

Fig 2.3 Similarity between one Documents

Avoiding duplicate work:

This algorithm determines the actually "Exact" duplicates and included in the list for the purpose of comparison of summary statistics for the full and fuzzy duplicates. As signature document used hash function MD5 [19] calculated for the entire document.

Ignore unimportant documents:

In this section we consider algorithms that use a training set of natural texts to build an automatic text generator. When the generation of words according to the formula (1) there are situations where there is no word from the nonzero probability of generation. This happens if at the end of the document met the unique sequence of k words. Choose an arbitrary state in the chain that starts causing. Then, at each step, select one of the possible transitions from the current state, with the generated word corresponding to this state. Spawn ends when reaches the desired length text.

III. PROPOSED METHOD

The level set method was in [13]. The topology changes of curves. A simple representation is that when a surface intersects with the zero planes to give the curve when this surface change, and the curve changes according to the surface changes.

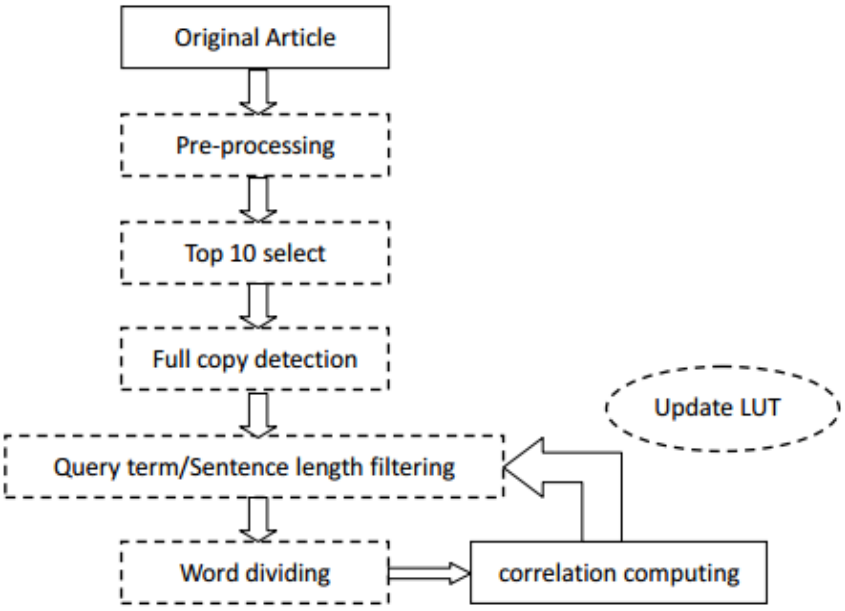

Fig 2.4 Look Up Table

The heart of the level set method is the implicit representation of the interface. To get an equation describing varying of the curve or the front with time, we started with the zero level set function at the front as follows:

$$\phi(x, y, t) = 0, \text{ if } (x, y) \in 1$$

Then computed its derivative which is also equal to zero

$$\frac{\partial \phi}{\partial t} + \frac{\partial \phi}{\partial x} \cdot \frac{\partial x}{\partial t} + \frac{\partial \phi}{\partial y} \cdot \frac{\partial y}{\partial t} = 0 \qquad (12)$$

Converting the terms of the dot product form of the gradient vector and the x and y derivatives vector, we go

$$\frac{\partial \phi}{\partial t} + \left(\frac{\partial \phi}{\partial x} \cdot \frac{\partial x}{\partial t} \right) \bullet \left(\frac{\partial \phi}{\partial y} \cdot \frac{\partial y}{\partial t} \right) = 0 \qquad (13)$$

Multiplying and dividing by $\nabla\phi$ and taking the other part to be F the equation was gotten as follows:

$$\frac{\partial\phi}{\partial t} + F|\nabla\phi| = 0 \tag{14}$$

According to literature [9][11], an energy function was defined:

$$E(\phi) = \mu E_{\text{int}}(\phi) + E_{ext}(\phi) \tag{15}$$

Where $E_{ext}(\phi)$ was called the external energy, and $E_{\text{int}}(\phi)$ was called the internal energy. These energy functions were represented as:

$$E_{\text{int}}(\phi) = \int_{\Omega} \frac{1}{2}(\nabla\phi - 1)^2 \, dxdy \tag{16}$$

$$E_{ext}(\phi) = \lambda L_g(\phi) + vA_g(\phi) \tag{17}$$

$$L_g = \int_{\Omega} g\delta(\phi)|\nabla\phi| dxdy \tag{18}$$

$$A_g = \int_{\Omega} gH(-\phi) dxdy \tag{19}$$

$$g = \frac{1}{1 + |\nabla G_\sigma * I|} \tag{20}$$

Where $L_g(\phi)$ was the lengthofff zero level curve of ϕ; and A_g could be viewed as the weighted area; another propagation method for the generation of unique texts is the union of different fragments of sample documents into one document. One variant of this algorithm is to partition the source code to the suggestions and compose a new text from an arbitrary sequence of statements. This method, as well as a method based on Markov chains, gives rise to the texts that have a local connection. Another important property that plays this method is the syntactic structure of natural Skye offers. Even a person cannot from a text - licit from natural, not getting a grasp on it. The main drawback of this approach - a higher probability of detection methods for analyzing such texts duplicates. In order to reflect the "circle " of the document in terms of the generator of texts, in the formula (3) is allowed zero or negative numbers document and such numbers counted backwards from the end of the construction of generalized chains on the example in Section. We use the formulas (3) and (4) and construct graphs of transitions between states of generalized circuits for different generators. It shows the transition graph generator based of length 1 trained on the documents d1 = {v1, v2, v3} and d2 = {v1, v1, v2}. The generator based on the fragments, trained on the same documents. All edges emanating from a vertex have the same weight (corresponding transition probabilities) and add up to one. Proposals document sample set of substrings of the text, the use of fingerprint, etc. 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. Therefore, recently a lot of attention is given to developing methods to reduce the computational complexity of algorithms created by selecting different heuristics. In applying the approximate approaches, decrease (sometimes very large) index completeness detects 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 "small" changes in the source documents and the ability to confidently handle short documents

IV. EXPERIMENT RESULTS

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 preserving the highest possible measure "accuracy". In this method, we experimented with complete sets of shingles and methods quadratic dependence, which occurs when a common feature has a large number of documents. The idea is based on (1) the observation that the size of words in the document is a good separating property and (2) by finding the total list of documents having the chain share the smaller, if the ratio of the lengths adjacent to at greater length exceeds a certain threshold, agrees the 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). Delete duplicates chains and chains that are entirely included in the other. As a result, the number of channels is reduced by hundreds of times, and the rest of the chain in the vast most are fairly short (2 - 10 items). The main disadvantage of the algorithm, reducing its performance, is the need to calculate the shingles for documents and use similarity function although, as previously stated, not all can be calculated shingles, but only some of them, according to a heuristic and replace more similarity with simple means. The below diagram exposes the F-measure that visualize the originality of multi documents.

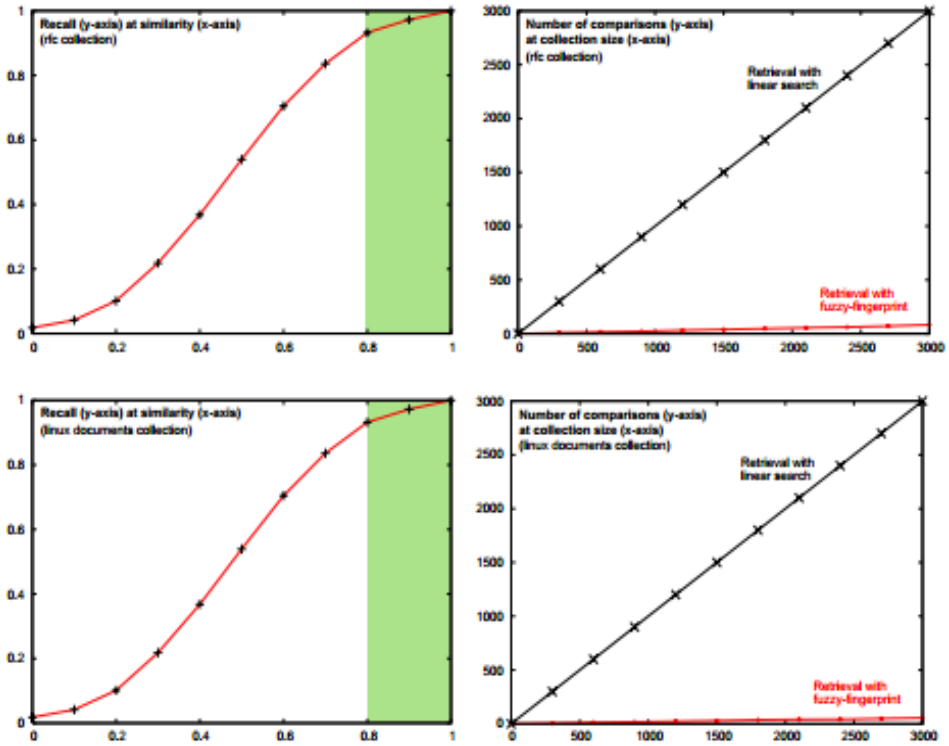

Fig 4.1F-measure: The plots on the left-hand side show the recall at similarity values that were achieved for the retrieval with fuzzy-document detection. The plots on the right-hand side illustrate the retrieval speed up and of the fuzzy-document detection retrieval process.

For the experiments, collection was divided into several parts comprising from three to twenty-four files (from about 5% to 50% of the entire size of the collection). Parameters used in the experiments: The number of words in the shingle 10 and 20, the spacing between adjacent beginning shingles 1. Indent This value means that the initial set of shingles includes all possible sequences word strings.

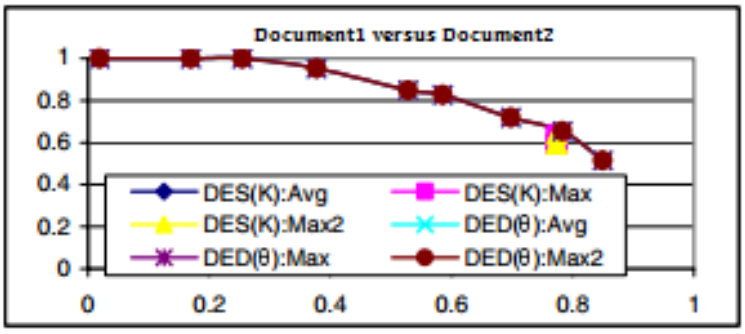

Fig 4.2 the fuzzy- cryptographic hashing on DES of average, and various maximum sequences.

In experiments with syntactic representations investigated by both methods of drawing up an image of the document elements as described in the permutation and " minimal elements in n permutations." Dimensions resulting document images were chosen, ranging from 100 to 200 shingles. In the case of " lexical " representation described in Section 2.1, for each document in the document data selected only words that fall into the final dictionary (a set of descriptive words). As a threshold determined by "frequent closed sets" (IE, the number of common shingles in data of documents from one cluster) we studied the different values in the intervals, the right of which coincides with the number of shingles in the form of a document, such as the interval [85 , 100] for data documents of 100 shingles interval [135 , 150] for data ddocuments of 150 shingles, etc. Obviously, when selected as the threshold value of the top slots in the cluster duplicates fell only those documents whose images coincide completely.

IV. DISCUSSION & CONCLUSION

For further research are presented is of interest approbation test based on special characters, different classes of documents. Provided the correct preliminary classification of the document possible tuning of particular set of parameters and weights of the algorithm creates the impression that may allow increased efficiency Vat duplicates in the particular class. According to the results of our experiments on the use of generating frequent closed sets in combination with conventional syntax and lexical means you can make the following conclusions. Methods of generating frequent closed sets are an effective way to determine the similarity of documents simultaneously with the generation of clusters of

similar documents. The results of syntactic methods duplicates considerable influence parameter "length of a shingle." Thus, in our experiments, the results for the length of the shingle of 10, were much closer to the top than doubles ROMIP shingle length of 20 , 15 and 5. Overall, the test results can be considered satisfactory. Low rates of low thresholds document similarity (compared, for example, the work of G. 'Softening' of the print can up - beat increase in the number of detecting pairs of duplicates, which will shift the efficient frontier at a low level. In this paper, we presented a new algorithm for finding duplicates on undistorted documents. A new method of calculating the hash based on the use of the LLP. The proposed solution showed the best results in the number of collisions compared to the existing solution to based on WHO. Recommendations on the choice of parameters of the detection algorithm - parameters coordinate strength pattern a, b, cardinality of the set. The method coordinate representation used template. Further research will be aimed at developing an algorithm based on the LLP under the field R Search for brightness- distorted duplicates of digital data. As part of this work Then, the developed and investigated generalized model unnatural texts generated by generators on the basis of the samples. The study illustrations that the texts generated by such generators, violate the thematic structure of natural texts. The algorithm uses a representation of the analyzed portions of the document as the valu\e of the specially selected hash function. Set of parameters selected on the basis of for reasons of stability to various kinds of modifications to the document. Thus, the experimented approach in terms of large volumes of documents and it is based on the theoretical model which was the proposed algorithm for detecting unnatural texts. It is on the analysis of the thematic structure of the texts.

REFERENCES

[1] Chow, T. W., & Rahman, M. K. M. (2009). Multilayer som with tree-structured data for efficient document retrieval and plagiarism detection. Neural Networks, IEEE Transactions on, 20(9), 1385-1402.

[2] Ananthakrishna, R., Chaudhuri, S., & Ganti, V. (2002, August). Eliminating fuzzy duplicates in data warehouses. In Proceedings of the 28th international conference on Very Large Data Bases (pp. 586-597). VLDB Endowment.

[3] Gamba P., & Savazzi, P. (1998, July). Classification of urban environments in SAR images: a fuzzy clustering perspective. In Geoscience and Remote Sensing Symposium Proceedings, 1998. IGARSS'98. 1998 IEEE International(Vol. 1, pp. 351-353). IEEE.. `

[4] Metwally, A., Agrawal, D., & El Abbadi, A. (2005, May). Duplicate detection in click streams. In Proceedings of the 14th international conference on World Wide Web (pp. 12-21). ACM.

[5] Weis, M., & Naumann, F. (2004, June). Detecting duplicate objects in XML documents. In Proceedings of the 2004 international workshop on Information quality in information systems (pp. 10-19). ACM.

[6] Leit Li, S., Son, S. H., & Stankovic, J. A. (2003, January). Event detection services using data service middleware in distributed sensor networks. In Information Processing in Sensor Networks (pp. 502-517). Springer Berlin Heidelberg.

[7] Koberstein, J., & Ng, Y. K. (2006). Using word clusters to detect similar web documents. In Knowledge Science, Engineering and Management (pp. 215-228). Springer Berlin Heidelberg.

[8] Aziz, A. M. (2011, March). A new fuzzy clustering approach for data association and track fusion in multisensor-multitarget environment. InAerospace Conference, 2011 IEEE (pp. 1-10). IEEE.

[9] Derr, E., & Glass, J. (2004). U.S. Patent Application 10/710,918.

[10] Puhlmann, S., Weis, M., & Naumann, F. (2006). XML duplicate detection using sorted neighborhoods. In Advances in Database Technology-EDBT 2006 (pp. 773-791). Springer Berlin Heidelberg.

[11] Christen, P. (2012). Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection. Springer.

[12] Ridley, M. J. (1992). An expert system for quality control and duplicate detection in bibliographic databases. Program: electronic library and information systems, 26 (1), 1-18.

[13] Draisbach, U., & Naumann, F. (2010). DuDe: The duplicate detection toolkit. In Proceedings of the International Workshop on Quality in Databases (QDB)

[14] Shahri, H. H., & Barforush, A. A. Z. (2004, June). Data mining for removing fuzzy duplicates using fuzzy inference. In Fuzzy Information, 2004. Processing NAFIPS'04. IEEE Annual Meeting of the (Vol. 1, pp. 419-424). IEEE.

[15] Grira, N., Crucianu, M., & Boujemaa, N. (2008). Active semi-supervised fuzzy clustering. Pattern Recognition, 41(5), 1834-1844.

[16] Lo, C. H., Chan, P. T., Wong, Y. K., Rad, A. B., & Cheung, K. L. (2007). Fuzzy-genetic algorithm for automatic fault detection in HVAC systems. Applied Soft Computing, 7(2), 554-560.

[17] Alzahrani, S., & Salim, N. (2009). Statement-Based Fuzzy-Set Information Retrieval versus Fingerprints Matching for Plagiarism Detection in Arabic Documents. In 5th Postgraduate Annual Research Seminar (PARS 2009), Johor Bahru, Malaysia (pp. 267-268).

