Adaptive Tag Based Document Explorer and Duplicate Detection Engine for Cloud

P.Basheer
M.Phil Scholar, Department of Computer Science, Sree Narayana Guru College, K. G. Chavadi Coimbatore, Tamil Nadu, India.

Abstract – Cloud is the large repository for different types of service. The cloud provides an effective way to store and share large size documents and files. The proposed work identifies security and privacy issues for cloud data storage using data mining techniques. An efficient privacy preserving data search and verification scheme is proposed to protect the user search privacy and content security on encrypted and uploaded cloud data. This also performs data redundancy detection on cipher text. Compared with the existing schemes, the proposed work only need to check a small portion of ranked indexes in a results and, thus, greatly reduces the verification cost. In the proposed system, explore a new technique named as "Adaptive Tag Based Fuzzy Clustering" is proposed for duplicate detection within the search result. In the proposed system, a bloom search method is proposed to support more search semantics and also to meet the demand for fast cipher text search within a dynamic huge data environment. The results shows, the engine provides high security and privacy for cloud data with increased search efficiency, accuracy and time efficiency.

Index Terms – Cloud Data Security, Duplicate Document, Data Search, Detection Approaches, Data Mining.

1. INTRODUCTION

In the recent trend, running business scenario, cloud data management is a big challenge. The high scalable cloud services[1][2] allows the user to store and share their documents over public and private clouds. This generates huge volume data [3]. With the huge amount of data from every user have several redundant and duplicate records. So duplicate data occupies more space and even increases the access time, this issue also creates several issues related to document search and security. Data mining is an effective way to solve such problems in the cloud service [4]. Data mining is the stepwise, possibly iterative and interactive process of extraction of unknown predictive information from large databases. It involves processing of huge volumes of data to discover patterns and trends that cannot be uncovered through normal database analysis. Data mining utilizes highly refined algorithms to process the data and to manage with large set.

Context analysis uses metadata associated with actual confidential data to find the duplication [5]. There are several approaches associated with the content duplication detection process. The duplicate files either created as a text document or sometimes other file formats. This redundancy can be due to several reasons. Duplicate detection techniques and methods is recently surveyed in [6] and found a set of issues. Lin, Yung-Shen al [7] authors have proposed a near-duplicate document detection technique that could be easily adjusted for a specific domain. Each important unique document has been represented as a real-valued sparse k-gram vector in their unique method. The weights have been trained to optimize for a particular important similarity function like cosine similarity or Jaccard coefficient process [8]. The enhanced similarity measure has been capable of reliably detecting the approximate-duplicate original documents. The efficient proposed similarity evaluation has been acquired by applying locality sensitive hashing scheme (SHS), which map these vectors to a few numbers of hash-values as important document signatures. After the deep analysis, the proposed system fuses the data mining and cloud security techniques to perform effective cloud data retrieval and duplication detection. In this paper, a new Adaptive Tag Based Fuzzy Clustering algorithm is proposed.

2. PROPOSED SYSTEM

This proposed work implements a new technique "Adaptive Tag Based Fuzzy Clustering" for duplicate file detection and leakage detection in the cloud data. Cloud data duplicate detection is one of the major tasks, which can reduce the storage intake and avoids redundant file retrieval. Cloud accompanying the rapid growth of data on the cloud and the growing need is to integrate data from heterogeneous sources for making the cloud storage less. In the cloud the documents are stored in the public storage system may have numerous duplicate data bear high similarity to each other, yet they are not bitwise identical. So the effective document duplication detection from the encrypted cloud storage is the main aim of the system.

The proposed system designed and developed a new Adaptive Tag Based Fuzzy Clustering (ATFC) for duplicate document detection in the cloud. The ATFC developed to bring the document management and document retrieval process over encrypted cloud data. In the proposed system, explore supporting different multi-keyword semantics over encrypted information and checking the integrity of the data within the search result. In the proposed system, ATFC method is proposed to support more search semantics and also to meet the demand for fast cipher text search within a dynamic huge data environment. The proposed system aims to provide security and privacy in terms of duplicate detection and fast data search for cloud data with increased search efficiency, accuracy and time efficiency. There are asset of algorithms are used to perform the de-duplication and data leakage in the cloud source. The followings are the contributions of the proposed system.

- The data encryption process for data security: this contains the set of processes such as (Setup, Encrypt, KeyGen, indexing and verification)
- ATFC is proposed to group the documents based on its similarity nature.
- Tag Entropy calculation between terms is performed for data grouping and analysis.

The proposed system developed an effective data mining framework integrated with the cloud storage system to detect the duplicate records and eliminating the data leakage. The proposed system utilized the data mining techniques to find the duplicate and near duplicate files in the cloud storage. The main advantage of this project is to find the duplicate files and eliminating the files from the cloud storage.

3. EXPERIMENT RESULT

As experimental data, the cloud data collection, consisting of files with the total size of 4.04 GB from cloud interface created for the experiment. For the experiments the collection is segregated by the cloud service providers, comprising from three to twenty four files (from about 5% to 50% of the entire collection).

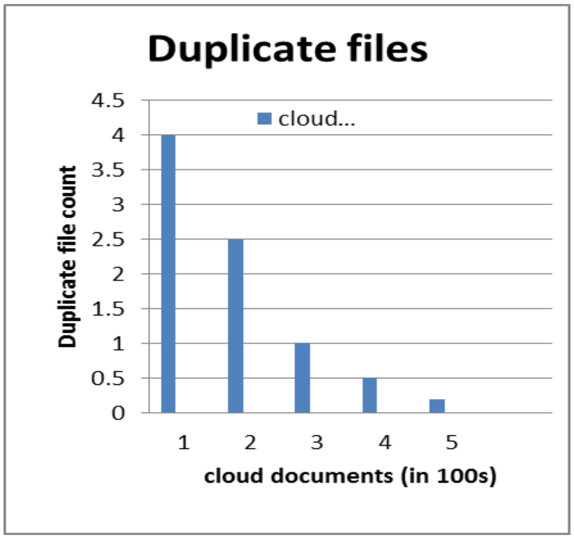

Figure 1.0: Identification of cloud file duplicate resources through adaptive fuzzy based method

Experiments show that the proposed method is fully automated for duplicate file detection and is a clustering algorithm that effectively combines applications. This can be designed for cloud storage as a component of its file management. Also, this could be a mechanism of independent research and development and improve the quality of the other cloud storage as a method, for example, search keywords. The first attempt would be to create an information system based on a set of files as a group. Search of local files is designed to search for files similar to the existing one from the encrypted content.

The investigation is to examine the input document and to show output files from the database that is most similar to the original. The implementation phase includes investigation of the existing system and its obstacles for careful planning and applying the design methods to achieve the transformation and evaluation methods of transformation. Implementation is the process of converting the theoretical design of the new system and bringing it into effect as a working system.

Fig 1.0 shows the identification of cloud file duplicate resources through data mining tag based method. One of the important characteristic of the duplicate document detection algorithms in the cloud is the ability to find the correct number of clusters. Robust metrics such as precision, recall and the F-measure which is widely used quality measures in cloud data retrieval and also suitable measures for evaluation of clustering algorithms in the task of duplicate cloud file detection are chosen.

In this method, the experimental results are evaluated, which were obtained through the evaluation metrics sensitivity, specificity and accuracy. In the performance metrics, the Sensitivity and specificity are the statistical measures and this is also known in statistics as clustering function. The Sensitivity will be calculated using the true positive rate and recall rate measures the proportion of actual positives which are correctly identified and is complementary to the false negative rate. Specificity is calculated from the true negative rate, and that is measured the proportion of negatives which are correctly identified, and is complementary to the false positive rate.

In order to find these metrics, the terms such as, “True positive”, “True negative”, “False negative” and “False positive” error on the basis of the definitions given below are first calculated. Files for a specific keyword can be distinguished between four disjoint sets and can be classified as combinations of the attributes “true” or “false” with “positive” or “negative”.

- True Positive = correctly identified
- False Positive = incorrectly identified
- True Negative = correctly rejected
- False Negative = incorrectly rejected

Table 1.0: Formula for evaluation metrics

Metrics	Formula
Precision (P)	P = TP / (TP + FP)
Recall (R)	R = TP / (TP + FN)
F-Measure (F_1)	$F_1 = (2\ \ P\ \ R) / (P + R)$
Sensitivity or True Positive Rate (TPR)	TPR = TP/P = TP / (TP + FN)
Specificity or True Negative Rate (SPC)	SPC = TN/N = TN / (FP+TN)
Accuracy (ACC)	ACC = (TP + TN) / (P+N)

The evaluation results in table 2.0, the values shows that the overall performance of the proposed approach provides encouraging results after applying Adaptive tag based Fuzzy clustering in terms of recall, precision and F-measure.

Table 2.0: Results of quality metrics for the Adaptive tag based Fuzzy technique

Quality Metrics	Proposed Approach without ATFC	Proposed Approach after applying ATFC
Recall (%)	85	88
Precision (%)	76	79
F-Measure (%)	80	83
False Positives (%)	28	29
False Negatives (%)	20	18

Table 3.0 Comparison of detection accuracy

Detection Accuracy (%)	***Models based Duplicate Document Detection***	***Near Duplicate file Detection***	***Proposed Fuzzy based Duplicate Document Detection***
	73.151	**90.687**	**96.852**

Although, as previously stated, each individual cannot be counted, but only some of them, according to the similarity heuristic and replace it with a simpler means. Algorithm showed nearly 100% completeness of the content has been detected duplication in a document based on a template of syntax. It includes details of the page, the text font and text captions, and the text within the table.

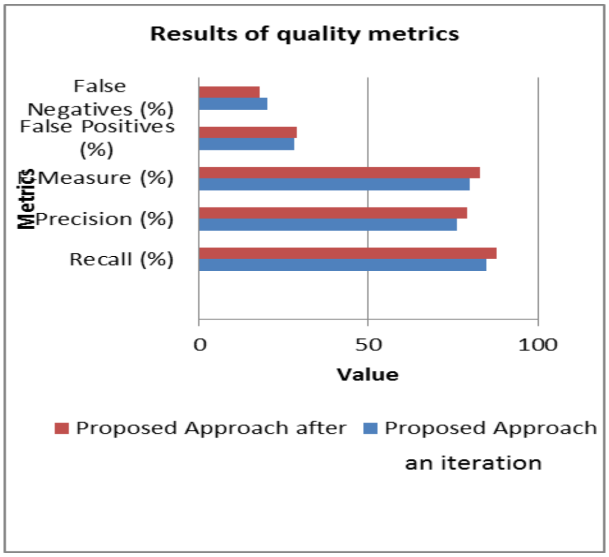

Figure 2.0: Evaluation metrics obtained for the proposed approach

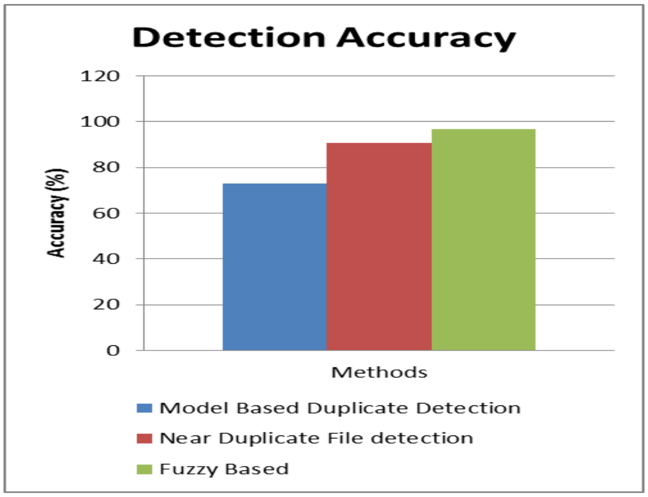

Figure 3.0 Detection accuracy

Upon completion of the search of a particular document, the algorithm does not compare the same data again to test for inclusion. It improves the efficiency of algorithm through the methods based on fuzzy. Construction of duplicates for the keyword is done by changing the line breaks and/or appending roughly certain amounts of unrelated text to the beginning and end of the document. From table 2.0 and figure 2.0, it can be observed that on the whole the performance of the proposed approach with Adaptive tag based Fuzzy clustering is improved in terms of high recall and precision and also low false positives and false negatives. The decrease in the number of false positives and false negatives shows that the proposed Adaptive tag based Fuzzy clustering accurately detects the duplicate files. The performance of the proposed approach with Adaptive tag based Fuzzy clustering is evaluated on the basis of accuracy on cloud dataset. The Adaptive tag based Fuzzy clustering has shown high F-measure of approximately 30% more than the approach without ATFC.

The detection accuracy is shown in figure 3.0 that the proposed Adaptive tag based Fuzzy algorithm is effective for the analysis of duplicate document in less time with reduced error in terms of clustering accuracy. The proposed method can also be used in the development of tools for the analysis of the dynamics of scientific and technical expertise in the collections of electronic files. In this section, a general method of algorithms is proposed based on duplicate document detection using fuzzy clustering. Through analysis of the existing methods of finding duplicate files, opportunities have not been exploited to increase the completeness of combined single system. Thus the aim is to overcome the disadvantages of existing algorithm in this regard. The algorithm is based on fuzzy cluster analysis of the similarity of different information in the document that contains all the pictures and the text available on the Internet through the application of the protocols. In the first phase, the algorithm assemble fuzzy duplicate content analysis of the image through RGB with the Euclidean distance metric, and in the second phase, is verified syntax similarity of text mining through the presentation of a border on the basis of time series sample for each document. Later, the similarity is verified in the analysis of texts through the presentation of a border on the basis of the time series for each sample document and compares the results with various existing approaches. Comparison results show the proposed algorithm is effective for the analysis of duplicate document in less time with reduced error in the clustering of other methods.

4. CONCLUSION

The focus of this research is to provide the data mining services to the detection of duplicate documents in the cloud encrypted source. This allows the user to search the content from the cloud over encrypted content based on the user requesting the information without duplication. Several methods have been proposed to detect duplicate and near duplicate files, but still needs some enhancement for the encrypted data reduplication. It is hypothesized that because fuzzy approach can measure the uncertainty between the class boundaries to identify the degree of duplication by membership values and because cloud encrypted documents contain complex and dynamic information for text clustering, Adaptive tag based Fuzzy clustering will be more effective with efficient tag selection, duplicate detection and document clustering. The major objective of this paper is to build up an enhanced duplicate document detection technique along with the leakage probability detection by clustering documents with high clustering accuracy, enhanced tag selection and also to decrease the convergence time and the number of iterations by duplicate detection and leakage prediction. The methodology used and developed for this research is the enhanced fuzzy techniques: Adaptive tag based Fuzzy clustering algorithm (ATFC) is proposed to perform duplication and data leakage. The upshot of the proposed work of each technique is addressed here. ATFC: This work focuses on proposals to the document sample set of substrings of the text, the use of tags, etc. In applying the approximate approaches, a decrease in index completeness detects duplicates and allows the search over encrypted data. In this approach, the standard syntactic and lexical approaches are used with different parameters as methods for submitting documents. The evaluation results shows that the overall performance of the proposed approach provides encouraging results after applying ATFC in terms of precision and F-measure values. In the future, this algorithm can be implemented to analyze mysterious text-based images to find similarities.

REFERENCES

[1] Kovatsch, Matthias, Martin Lanter, and Zach Shelby. "Californium: Scalable cloud services for the internet of things with coap." *Internet of Things (IOT), 2014 International Conference on the*. IEEE, 2014.

[2] Buyya, Rajkumar, Rajiv Ranjan, and Rodrigo N. Calheiros. "Modeling and simulation of scalable Cloud computing environments and the CloudSim toolkit: Challenges and opportunities." *High Performance Computing & Simulation, 2009. HPCS'09. International Conference on*. IEEE, 2009.

[3] Bhadani, Abhay Kumar, and Dhanya Jothimani. "Big Data: Challenges, Opportunities, and Realities." *Effective Big Data Management and Opportunities for Implementation, IGI Global, Pennsylvania, USA* (2016): 1-24.

[4] Dillon, Tharam, Chen Wu, and Elizabeth Chang. "Cloud computing: issues and challenges." *Advanced Information Networking and Applications (AINA), 2010 24th IEEE International Conference on*. Ieee, 2010.

[5] Piernas, Juan, Toni Cortes, and José M. García. "DualFS: a new journaling file system without meta-data duplication." *Proceedings of the 16th international conference on Supercomputing*. ACM, 2002.

[6] Sivakumar, T., and P. Basheer. "A Survey on Data Leakage Detection and De-Duplication in Data Mining System." *Journal of Network Communications and Emerging Technologies (JNCET) www. jncet. org* 7.10 (2017).

[7] Lin, Yung-Shen, Jung-Yi Jiang, and Shie-Jue Lee. "A similarity measure for text classification and clustering." *IEEE transactions on knowledge and data engineering* 26.7 (2014): 1575-1590.

[8] Niwattanakul, Suphakit, et al. "Using of Jaccard coefficient for keywords similarity." *Proceedings of the International MultiConference of Engineers and Computer Scientists*. Vol. 1. No. 6. 2013.