A Survey on Data Leakage Detection and De-Duplication in Data Mining System

T.Sivakumar
Assistant Professor, Department of Computer Science, Sree Narayana Guru College, K. G. Chavadi, Coimbatore, Tamil Nadu, India.

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

Abstract – In the recent trend, running business scenario, data leakage is a big challenge as critical organizational. In many organizations or agencies, this private organizational data have been shared to outside the organizational premises. When the distributor's sensitive data have been leaked by agents, and is more difficult to identify the agent that leaked the data. With the huge amount of data from every distributor's may be Having duplicate records in amount of data .So duplicate data occupies more space and even increases the access time ,Also it creates several issues relating to search and security. Data mining is an effective way to solve such problems in the cloud service. This paper surveys various techniques and methods used to is to detect an agent who leaks any portion of the owner's data and detect duplicate records in the cloud storage service.

Index Terms – Data leakage, Duplicate document, guilty agent, Leakage detection, Detection approaches, data mining.

1. INTRODUCTION

Cloud Computing has become a main source for the data processing, storage and distribution infrastructure that provides resources and/or services over the Internet. The storage of the data is simple and free to use. In data mining the data which is used as data security in a cloud computing platform. Cloud computing offers a new way of Information Technology services by rearranging various resources (e.g., storage, computing) and providing them to users based on their demands [1]. Cloud computing provides a big resource pool by linking network resources together. It has desirable properties, such as scalability, elasticity, fault tolerance, and pay-peruse. This allows it to maintain these data. Cloud computing security is an advancing there are several challenges associated with data prevention, data authorization and data leakage detection.

Figure 1.0 shows the general structure of cloud computing with the Data Leakage detection, File duplicate detection theme. Cloud is an alternative way for providing online services which are basically on demand such as networks, storage, server, software. It saves hardware cost and time by allowing pay by perusing i.e. pay according to user usage. In other words, Cloud computing is a mixture of new technology and platforms that provide storage and hosting services on the internet .The cloud technology is totally dependent on the internet where the data is stored in its data centers of the service providers. By using any device like PDA (Personal Digital Assistant), mobile etc can able to access the services. Cloud-based services for large scale content storage, processing, and distribution. Security and privacy are among top concerns for the public cloud environments [2]. Detection or avoidance of data leakage and misuse is a great challenging issue for organizations. Whether caused by the malicious intent or an inadvertent mistake, data leakage and misuse can damage the reputation. This challenge becomes more difficult when trying to detect and/or prevent data leakage and misuse performed by an insider having legitimate permissions to access the organization's systems and its sensitive data [3]. De duplication task is to find a function that can resolve when two records refer to the same entity in spite of errors and inconsistencies in the data. De duplication is a task of identifying record replicas in a data repository that refer to the same real world entity or object and systematically substitutes the reference pointers for the redundant blocks; also known as storage capacity optimization the amount of repeated data takes a toll on storage availability [4].

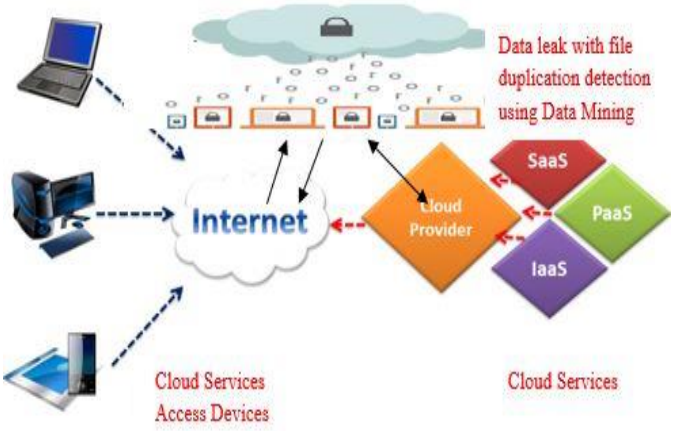

Fig 1.0 Cloud Infrastructures

There are several issues threatens the cloud environment by above issues at the time of data store in cloud services. to the protect data from leakers traditionally, Data leakage happens every day when confidential information are leaked out. This uncontrolled data leakage put serious risk for user. Now individuals as well as the corporate companies are moving to the cloud and as such data leakages have become a challenge task for leakage detection is handled Watermarks can be very useful in some cases, the problem is that these watermarks can be sometimes destroyed if the recipient is malicious. Data mining techniques and application are very much needed in the cloud computing paradigm for privacy protection, such ranking operation and data search may need access privilege to do that, however, some operations are allowed in this case, it should not assure that the data can be completely protected. To improve search result accuracy as well as to enhance the user searching experience, it is also necessary for data mining techniques to support keyword searches on both encrypted and non-encrypted contents. And the techniques should perform a keyword search and effective indexing with the privacy of data and efficient searching schemes, real privacy is obtained only if the user's identity remains hidden from the Cloud Service Provider as well as the third party user on the cloud server. In this paper, the secure and effective Data Leakage Detection and de-duplication techniques are reviewed.

A. Data Leakage Detection Techniques

Data leakage detection is handled by watermarking Techniques [5] using this approach a unique code is embedded in each distributed copy. If that copy is later discovered in the hands of an unauthorized party side, the leaker can be identified. Watermarks can be very useful in some cases only but again it's involving some modification of the original watermarks can sometimes be destroyed if the data recipient is attacker. Data leakage is a silent but destructive type of threat in data sharing application Sensitive information can be leaked without any knowledge. In this paper [6] Author is used to Invisible Watermarking technique presents a novel invisible robust watermarking scheme for embedding and extracting a digital watermark in to an image. Invisible insertion of the watermark is performed in the most significant region of the host image this invisible watermark signal is embedded in such a way that it cannot be removed without affecting the quality and efficiency of the data. This Invisible Watermarking instance be used for copyright protection as it can hide information about the author details in the data. The watermark can now be used in to prove ownership in court. Some of the approach conducts in this paper A DLP classification model was proposed based on the well-known information retrieval function. The classification was based on measuring the similarity between the documents and the category. This model was tested against different scenarios with known and unknown data, and partially known. Finally the overall classification shows and encouraging outcomes across all scenarios [7]. Data Leakage Prevention System with Time Stamp: In this paper, document along with the data one more parameter i.e. time stamp also considered as an important aspect in the Data leakage Prevention. In this approach Learning Phase the documents are trained confidential documents with time stamp. After that In the detection phase the tested document is compared with confidential score and time stamp, if the time stamp of the tested document is greater than or equal to the time stamp in the above table then that document is treated as a confidential and it is blocked [8]. In this paper [9] author proposed Agent Guilt Model technique proposed The agent who involved in irresponsible activity can be detected by the probability value. The probability of agents who are not authorized to receive data to the total number of agents for whom the particular data had been sent. Agent Guilt Model is used to assess the likelihood that leaked data came from one or more agents, by finding out the probability of value depicts the agents who are guilty. To compute this data leaker, this system needs estimation for the probability that values in dataset which can be "guessed" by the target. This model supports the identification of guilty agent based on the rights which are given to the agents [10].

Data allocation strategies technique invokes identifying leakages .These techniques not do any alterations of the released data. Sometimes also inject realistic but fake data records to the further improve our chances of detecting leakage and identifying of the guilty agent in accurate manner. The data allocation problem is focused in how the distributor can order to give the data intelligently to the agents by the distributor in order to improve the chances of detecting a guilty agent [11]. Identify the guilty agent address this based on the requests and the fake objects insertion also present algorithm for distributing object to agent. In paper [12] authors used is to detect when the distributor's sensitive data has been leaked by agents and identify the agent who has leaked the data using data allocation strategies. Some of the approaches are find out identify the data leakage In paper present Guilt of agents this model distributing objects to different agents, in a way that improves our chances of identifying a data leaker. Finally also consider the option of adding the fake objects into the distributed set of data given by distributor part. Such these objects do not correspond to real entities but appear realistic to the agents for this add fake object. Hence the fake objects act as a type of watermark for the given entire set, without modifying any individual agents. Suppose if it finds that an agent was given one or more fake objects that were leaked then the distributor of system the distributor has share some objects with different agents so it is likely to suspect them, leaking the data can be more confident that agent was guilty. In the System the data leaker can be traced with good amount of evidence [13].In this paper [14] Author proposed Privacy-preserving mining with data leakage detection with the help of data streaming model (from which association rules can easily be computed) on an

encrypted outsourced Transactional Data Base (TDB). This assumed that a conservative model where the adversary knows the domain of items and their exact frequency and can use this knowledge to identify cipher items and cipher item sets. It also presents algorithms for distributing objects to agents, in a way that improves our chances of identifying a leaker.

B. Data Duplicate Detection Techniques:

The data duplication is one of the critical issues in the data mining. Many industries will look for the accurate data to carry out their operations. Therefore the data quality must be significant. With the increase in the volume of data even the data quality problems arise. There are several approaches for solving duplicate detection problem Progressive Sorted Neighborhood Method (PSNM) depends on the conventional Sorted Neighborhood Method: PSNM sorts the information utilizing a predefined sorting key and just thinks about records that are inside the window of records in the sorted request. The instinct is that records that are close in the sorted request will probably be copies than records that are far separated, on the grounds that they are as of now comparative as for their sorting key. All the more specifically, the separation of two records in their short positions (rank-separation) gives PSNM an evaluation of their coordinating probability [15]. The PSNM calculation utilizes this instinct to iteratively change the window size, beginning with a little window of size two that rapidly finds the most encouraging records. This performs best on small and almost clean datasets. PSNM sorts the input data victimize a predefined sorting key and only compares records that are within a window of records in the sorted order it is performed over clean and small dataset [16].

Parallel Progressive Sorted Neighborhood Method PPSNM:- Progressive duplicate detection algorithms apply on selective input dataset (Cluster) that significantly increase the efficiency of finding duplicates if the execution time is limited. Duplicate detection is done on this phase .PSNM detect duplicate records sequence time manner So that Execution Time is higher than PSNM .Progressive Sorted Neighborhood Method used for Detecting Duplicate Records in within the minimum amount of time as compare with simple Sorted Neighborhood Method [17]. The main drawback of PSNM is Time Complexity because it detecting records duplicate in serial wise. This approach used to find duplicate detection with parallel approach, enhance Map Reduce algorithm for limit the execution time interval [18].

Duplicate Count Strategy ++ The Duplicate Count Strategy (DCS++) gets over fixed size window and introduces adaptive windows that vary size on identified duplicate within that window without affecting the efficiency and effectiveness of SNM. DCS++ starts with a domain dependent initial window of size just like SNM.This method is an extension of DCS Duplicate Count Strategy information. It is a strategy which dynamically adapts the window size it varies the window size based on the number of duplicates records detected. This is based is based on the intuition that there might be regions of high similarity which requires a larger window size and regions of lower similarity which requires a smaller window size information.Adaptation will sometimes increase or decrease the number of comparisons, if more duplicates of a record are found within a given window. The larger the window should be if no such duplicate of a record within its neighborhood is found, assume that there are no duplicates or the duplicates are very far away in the sorting order [19].In this paper [20] proposed PSO Algorithm Based Deduplication proposed a heuristic global optimization method called Particle Swarm Optimization algorithm for record deduplication. They considered the fitness function of the PSO algorithm and it is based on swarm of data. Here the proposed approach has two phases such as training phase and duplicate detection phase. First they find the similarity between the all attributes of record PSO algorithm is used to generate the optimal similarity measure for the training datasets. Once the optimal similarity measure obtained, the deduplication of remaining datasets is done with the help of optimal similarity measure generated from the PSO algorithm. PSO algorithm is very simple and it needs This algorithm has no overlapping and mutation calculation. It provides more accuracy in record deduplication than genetic algorithm. Due to the duplicate of records and dirty data, many problems will occur like performance degradation, quality loss and increasing operational data costs [21]. In this paper [22] proposed Divide and Conquer Based De duplication suggested an approach for duplicate record detection and removal. In this approach, they first convert the attributes of data into numeric form of information. Then, this numeric form is used to create clusters by using K-Means data mining clustering algorithm. The use of clustering reduces the number of comparisons operation of total dataset. After that the divide and conquer technique is used in parallel with these clusters for identification and removal of duplicated records in proper manner. Here, this technique identifies all type of duplicated records like fully duplicated records, erroneous duplicated records and also partially duplicated records. This technique is only applicable for single table instead of multiple sorted tables that's the major problem with this technique. Even though the existing duplication techniques are more beneficial and cost effective, but it suffers from several security vulnerabilities. So detection of document similarity and document search need a security consideration too. Out of line deduplication In that technique when data is stored into the server then deduplication task is performed. That reduces the possibility of loss of data. In that a large space is required first to store data. That generates space overhead for the technique. Inline Deduplication In that technique a deduplication task is performed at client end and then data transmitted over the network channel. That reduces the space over head for the storage. In that, an enhanced technique is required to provide

better performance to deduplicate data. It generate network bottleneck problem [23].

C. Deduplication over encrypted Data

In order to provide data security and privacy in cloud data storage service, users encrypts their data before uploading. This process may thwart security issues in cloud server and outside attackers. However, conventional encryption under different users' keys makes cross-user de-duplication unfeasible due to the different cipher texts for the same data.

Recently, authors in [24] proposed a convergent key management scheme. Using this scheme the users delivers the convergent key shares across several servers. This has been made with the Ramp secret sharing scheme. An authorized de-duplication scheme in which differential privileges of users, as well as the data, are considered in the de-duplication procedure in a hybrid cloud environment is proposed in [25].

Authors in [29] projected an anonymous de-duplication scheme over encrypted data that exploits a proxy re-encryption algorithm.

Later a server-aided MLE is proposed in [26]. This is secure against brute-force attack and other data misuse attacks. The work has recently extended to interactive MLE to provide privacy for messages that are both correlated and dependent on the public system parameters. But, these schemes do not handle the dynamic ownership management issues involved in secure de-duplication for shared outsourced data.

Authors in [27] proposed a de-duplication scheme over encrypted data that uses predicate encryption. This approach allows de-duplication only of files that belong to the same user, which severely reduces the effect of de-duplication. Thus, the proposed work focused on de-duplication across different users such that identical files from different users are detected and de-duplicated safely to provide more storage savings.

2. PROBLEM STATEMENT

One challenge to detect when the distributor's sensitive data have been leaked by agents along with identification of Data Deduplication. Deduplication is most effective when multiple users outsource the same data to the cloud storage, but it raises issues relating to security and ownership security. Security requirement to protect against not only outside adversaries but also inside the cloud server .One of the challenging issue in secure deduplication over encrypted data and identify data leakage detection in cloud storage.

- The existing techniques is failed to perform document leakage detection because of these watermarks can be sometimes destroyed if the recipient is malicious.
- The Agent Guilt model completely not suitable for complete privacy.
- The searching accuracy in fully encrypted dynamic data is low
- Computational overheads are high
- Existing approach increasing the number of comparison between the records such that it increases the time consumption

However, applying the data leakage detection on the data and user information in the encrypted cloud data search system remains a very challenging task because of inherent security and privacy obstacles, including various strict requirements like the data privacy, the search privacy, the tag generation privacy, and many others.

PID	Technique	Advantages	disadvantages
14	sorted neighborhood method	Compare all records within the window	some duplicates might be missed
15	Progressive Sorted Neighborhood Method	Data duplication detection is fast	decrease the efficiency of the duplicate detection
17	Parallel Progressive Sorted Neighborhood	detection of duplicates is very easy because the comparison among all possible duplicate pairs is no required	performed over perfect and little datasets
19	Duplicate Count Strategy ++	Find duplicates in the large dataset	Couldn't detect for new record
20	Particle Swarm Optimization algorithm	Fast	it is not possible to assume a unifying set of standards for various data sources

25	Ramp secret sharing scheme	Optimized for data integration and analytical process	Require accurate test samples
26	BL-MLE	Used to find duplicate entries in the book domain	Not suitable for cloud

Table 1.0 comparative study of the data leakage detection and de-duplication techniques in data mining

Data leakage detection techniques can be categorized into different challenges associated with data prevention, data authorization and data leakage detection. In the former approach, most of the existing schemes have been proposed in order to perform Data leakage detection and data prevention process in an efficient and robust manner, since this approach performs prediction and probability finding in order to identify and protect data from leakers, is vulnerable to being leaked to outside adversaries because of its prediction and probability cannot perform efficient manner. In the latter approach, data privacy is the primary security requirement to protect against not only outside adversaries but also inside the cloud server. Thus, most of the schemes have been proposed to provide data encryption, while still benefiting from a de-duplication technique, by enabling data owners to share the encryption keys in the presence of the inside and outside adversaries. Since encrypted data are given to a user, data access control can be additionally implemented by selective key process.

3. OBJECTIVE FOR THE FUTURE WORK

From the survey, some objective of the future work is identified and gathered. The objective of the further research is to improve the accuracy of finding data leakage detection and duplicate record detection process in the cloud storage services. A domain independent approach is carried out to detect the duplicate records available in the large databases. Providing a cost effective non cryptographic anonymity protection against privacy data sharing and fuzzy based anomaly detection. Extended k anonymity model this can design more effective grouping algorithms to ensure better protection against data disclosure and solution against data leakage attacks.It makes use of data mining similarity functions in detecting the duplicate contents. Along with the clustering method additional indexing techniques can be used to reduce the time taken on each comparison to improve duplicate detection.

Finally, the work addresses the problem of protecting owner data leakage information with to ensure better protection against data disclosure developing extended k anonymity model and solution against data leakage attacks threshold definition for similarity measures and tag definition of cloud data search; this can be expanded by automatically generating the tags and thresholds which achieves more accuracy besides reducing errors. The work obtained from the existing scheme provides the following improvement ideas such as; it should improve the accuracy of duplicate record detection process, it should reduce the time taken to detect the duplicate using clustering, it should find the optimized expression which shows weightage of the attributes that plays an important role in identifying the duplicates and finally, a complete and effective indexing methods should be used for fast retrieval.

4. CONCLUSION

In this paper, the problem of finding of data leak detection and eliminating detection duplicate records/document using data mining techniques are investigated. The efficient identification of duplicate records in the distributed system is a vital issue that has occurred from the increasing amount of data and the necessity to integrate data from diverse sources and needs to be enhanced. In this paper, a comprehensive survey of researches of data leak detection and Duplicate document detection and de-duplication techniques using data mining in cloud storage services is proposed. The review summarizes, that there is no enough study carried out to handle de-duplication, data leak detection, efficient search over encrypted content is deployed for cloud storage services. Because, the current trend is fully based on the cloud, so effective cloud data management is necessary with optimal data de-duplication and data leak detection.

REFERENCES

[1] Wu, Jiyi, et al. "Cloud storage as the infrastructure of cloud computing." *Intelligent Computing and Cognitive Informatics (ICICCI), 2010 International Conference on*. IEEE, 2010.

[2] Wu, Jiyi, Lingdi Ping, Xiaoping Ge, Ya Wang, and Jianqing Fu. "Cloud storage as the infrastructure of cloud computing." In *Intelligent Computing and Cognitive Informatics (ICICCI), 2010 International Conference on*, pp. 380-383. IEEE, 2010.

[3] Kumar, Neeraj, et al. "Detection of data leakage in cloud computing environment." *Computational Intelligence and Communication Networks (CICN), 2014 International Conference on*. IEEE, 2014.

[4] Low, Wai Lup, Mong Li Lee, and Tok Wang Ling. "A knowledge-based approach for duplicate elimination in data cleaning." *Information Systems*26.8 (2001): 585-606.

[5] Kumar N, Katta V, Mishra H, Garg H. Detection of data leakage in cloud computing environment. InComputational Intelligence and Communication Networks (CICN), 2014 International Conference on 2014 Nov 14 (pp. 803-807). IEEE.

[6] Han, Yanyan, et al. "A digital watermarking algorithm of color image based on visual cryptography and discrete cosine transform." P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC), 2014 Ninth International Conference on. IEEE, 2014.

[7] Alneyadi, Sultan, Elankayer Sithirasenan, and Vallipuram Muthukkumarasamy. "Adaptable n-gram classification model for data leakage prevention." *Signal Processing and Communication Systems (ICSPCS), 2013 7th International Conference on*. IEEE, 2013.

[8] Peneti, Subhashini, and B. Padmaja Rani. "Data leakage prevention system with time stamp." *Information Communication and Embedded Systems (ICICES), 2016 International Conference on*. IEEE, 2016.

[9] Kumar, Neeraj, et al. "Detection of data leakage in cloud computing environment." *Computational Intelligence and Communication Networks (CICN), 2014 International Conference on*. IEEE, 2014.

[10] Dhanalakshmi, V., and R. Shagana. "Assess agent guilt model and handling data allocation strategies for data distribution." In *Intelligent Computing and Cognitive Informatics ,2013 International Conference on*, pp. 1-5. IEEE, 2013.

[11] Chen, Tzung-Shi, and Jang-Ping Sheu. "Communication-free data allocation techniques for parallelizing cloud storage on multiVM." *IEEE Transactions on knowledge and data engineering* 19.1 (2007): 1-16.

[12] Elmagarmid, Ahmed K., Panagiotis G. Ipeirotis, and Vassilios S. Verykios. "Data Leakage detection: A survey." *IEEE Transactions on knowledge and data engineering* 19.1 (2007): 1-16.

[13] Ahmad, Miss SW, and G. R. Bamnote. "Data leakage detection and data prevention using algorithm." *international conference on Management of data* 6.2 (2013).

[14] Shu, Xiaokui, Danfeng Yao, and Elisa Bertino. "Privacy-preserving detection of sensitive data exposure." *IEEE transactions on information forensics and security* 10.5 (2015): 1092-1103.

[15] Thampi, Shanila, and D. Loganathan. "Progressive of Duplicate Detection Using Adaptive Window Technique."

[16] Elmagarmid, Ahmed K., Panagiotis G. Ipeirotis, and Vassilios S. Verykios. "Sorting key Based Duplicate Detection Using Adaptive Window Technique: A survey." *IEEE Transactions on knowledge and data engineering* 19.1 (2007): 1-16.

[17] Dhane, Shubhangi A., and Amrit Priyadarshi. "Survey On Parallel Duplicate Detection."

[18] Cochinwala, Munir; Verghese Kurien and Gail Lalk and Dennis Shasha (2001). "Efficient Parallel Duplicate Data Detection ". Information Sciences 137 (1-4): 1-15.

[19] Skandar, Arfa, Mariam Rehman, and Maria Anjum. "An Efficient Duplication Record Detection Algorithm for Data Cleansing." *International Journal of Computer Applications* 127.6 (2015): 28-37.

[20] Sudhakaran, Saniya, and Meera Treesa Mathews. "A Survey on Data De-duplication in Large Scale Data using PSO." *International Journal of Computer Applications* 165.1 (2017).

[21] Chiang, Yueh-Hsuan, AnHai Doan, and Jeffrey F. Naughton. "Modeling entity evolution for temporal record matching suing K means clustring" *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*. ACM, 2014.

[22] Randall, Sean M., et al. "The effect of data cleaning Divide and Conquer Based De duplication suggested." *BMC medical informatics and decision making* 13.1 (2013): 64.

[23] Hamid HaidarianShahri, Saied HaidarianShahri, "Eliminating Duplicates in information using Divide and Conquer Integration: An Adaptive, Extensible Framework", IEEE Computer Society 1541-1672, pp. 63-71, September/October 2006.

[24] L. Padmasree, V. Ambati, J. Chandulal, and M. Rao. Signature Based Duplication Detection in Digital Libraries. Signature, 2006.

[25] Chaudhuri, Surajit, Venkatesh Ganti, and Raghav Kaushik. "A primitive operator for similarity joins in data cleaning." *Data Engineering, 2006. ICDE'06. proceedings of the 22nd International Conference on*. IEEE, 2006.

[26] Yan, Su, et al. "Adaptive sorted neighborhood methods for efficient record linkage." *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*. ACM, 2007.

[27] Bayardo, Roberto J., Yiming Ma, and Ramakrishnan Srikant. "Scaling up all pairs similarity search." *Proceedings of the 16th international conference on World Wide Web*. ACM, 2007.

[28] Tateishi, Kenji. "Dai Kusui. Fast Duplicate Document Detection using Multi-level Prefix-filter." *The Third International Joint Conference on Natural Language Processing*. 2008.

[29] Paskalev, Plamen, and Anatoliy Antonov. "Increasing the performance of an application for duplication detection." *Proceedings of the 2007 international conference on Computer systems and technologies*. ACM, 2007.

[30] Culotta, Aron, and Andrew McCallum. "Joint deduplication of multiple record types in relational data." *Proceedings of the 14th ACM international conference on Information and knowledge management*. ACM, 2005.

[31] Chum, Ondrej, James Philbin, and Andrew Zisserman. "Near Duplicate Image Detection: min-Hash and tf-idf Weighting." *BMVC*. Vol. 810. 2008.

[32] Beskales, George, Mohamed A. Soliman, and Ihab F. Ilyas. *Modeling Uncertainty in Duplicate Elimination*. Technical Report, March 31, 2008.

[33] Christen, Peter. "Automatic record linkage using seeded nearest neighbour and support vector machine classification." *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2008.

[34] Elhadi, Mohamed, and Amjad Al-Tobi. "Duplicate detection in documents and webpages using improved longest common subsequence and documents syntactical structures." *2009 Fourth International Conference on Computer Sciences and Convergence Information Technology*. 2009.

[35] D. Harnik, B. Pinkas, and A. Shulman-Peleg, "Side channels in cloud services, the case of deduplication in cloud storage," IEEE Security & Privacy, vol. 8, no. 6, pp. 40–47, 2010.

[36] S. Halevi, D, Harnik, B. Pinkas, and A. Shulman-Peleg, "Proofs of ownership in remote storage systems," Proc. ACM Conference on Computer and Communications Security, pp. 491–500, 2011.

[37] M. Mulazzani, S. Schrittwieser, M. Leithner, and M. Huber, "Dark clouds on the horizon: using cloud storage as attack vector and online slack space," Proc. USENIX Conference on Security, 2011.

[38] J. Li, X. Chen, M. Li, J. Li, P. Lee, and W. Lou, "Secure deduplication with efficient and reliable convergent key management," IEEE Transactions on Parallel and Distributed Sytems, Vol. 25, No. 6, 2014.