ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

A Survey on Data De-Duplication Methods in Cloud Storage System

S. Lalitha [1], N. Kamal Raj [2]

M.Phil Scholar, Department of Computer Science, Dr.SNS Rajalakshmi College of Arts and Science, Coimbatore, Tamil Nadu, India[1]

HOD, Department of Information Technology, Dr.SNS Rajalakshmi College of Arts and Science, Coimbatore, Tamil Nadu, India[2]

ABSTRACT: In the recent trend, every data and contents are stored in the cloud using cloud storage services. With the huge amount of data from every client may affect the cloud storage. In specific, the redundant content may perform more worst in the storage part. The de-duplication method is generally used to reduce the storage cost and resource requirements of data services in the cloud by eliminating redundant data and storing only a single copy of them. De-duplication is most effective when multiple users outsource the same data to the cloud storage services, but 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 detect duplicate records in the cloud storage service.

KEYWORDS: Cloud storage, Duplicate document, near duplicate pages, near duplicate detection, Detection approaches, data mining.

I. INTRODUCTION

Cloud computing is growing in terms of powerful resources and types of computing services from the last few years. It is a method of delivering technology to the consumer. Cloud computing provides dynamic, scalable and pay-as-you-go service models which can be adopted by many organizations to save their expenditure and time to build required IT infrastructure [1]. The cloud computing is another process computing, distributed computing, parallel computing, virtualization technology, utility computing and other computer technologies. For example, large scale computation and data storage, virtualization, high expandability high reliability and low price service. The security issue of cloud computing is extremely paramount and it can keep the fast improvement of cloud computing. Cloud computing security is an advancing sub-area of network security.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

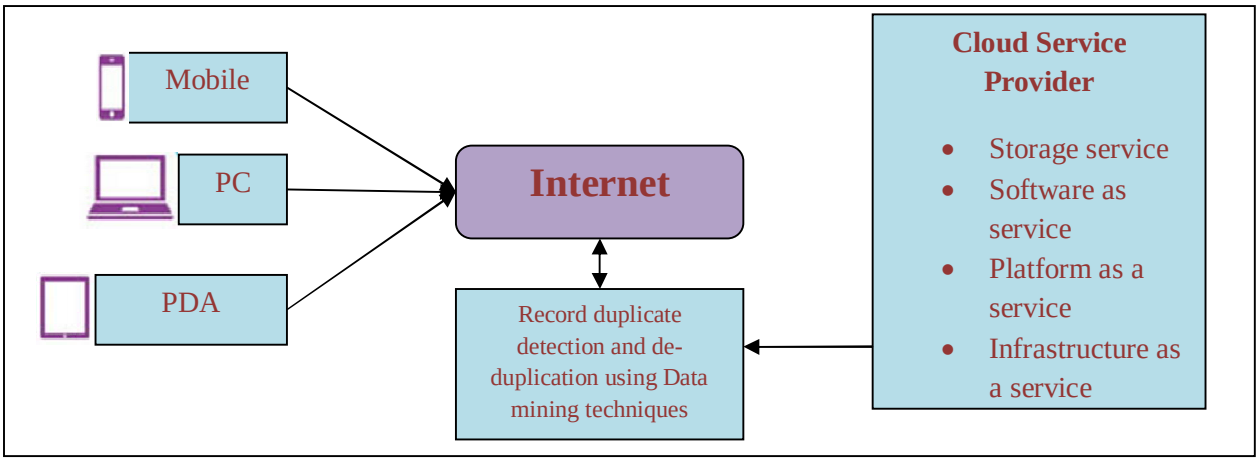

Fig 1.0 Cloud Infrastructures

Figure 1.0 shows the general structure of cloud computing with the de-duplication detection theme. Using any device like PDA (Personal Digital Assistant), mobile etc, the cloud can be accessed through Cloud Service Provider (CSP). The cloud computing architecture consists of components of f platform, back end platform and a network host, for example, PDA, mobile etc., computers, servers, and data storage services. Storage clouds play a significant role in distributed systems. A number of organizations require high data quality with least overhead. Data quality problems arise with the constantly increasing quantity of data stored in real-world databases that are assured by the vital data clean-up process [2]. Data quality problems are encountered in the single data collections, like the files and databases. Data pre-processing techniques deals with the detection and removal of redundant and error contained data and inconsistencies from the data to improve the quality of data in the cloud storage service. It is necessary to enhance the quality of data in a cloud data server. Numerous data cleaning techniques are being employed for diverse purposes. The fundamental element of data cleaning is usually termed as duplicate content identification that is the process of identifying the record pairs signifying the same records. The process of duplicate detection is preceded by a data preparation stage, which includes an indexing, data analysis, cloud security privilege verification and data standardization etc., Data de-duplication in Cloud storage service encompasses many technologies [3]. And this also has several security considerations like privileges to Access servers and applications, Virtual Machine Security, Data Privacy, Data Integrity, Data Security and Security policy and compliance.

There are several security issues threatens the cloud environment by above issues at the time of data retrieval and de-duplication. To protect the privacy of data and resist unwanted accesses in the cloud storage data. Because the storage data can be a sensitive data, which categorized in a different format like health data, personal photos, secure tax contents etc., so, theses contents should be encrypted by data owners before outsourcing to the cloud storage. While considering the content retrieval and data search, the traditional service is based on plain-text keyword search methods. The irrelevant process of downloading all the data and decrypting locally is clearly impractical and not cost effective. Due to a large amount of bandwidth cost in cloud scale systems, it creates a big trouble. The uploaded contents are sometimes treated as important information for the other category, Therefore, exploring privacy preserving and effective search service over encrypted cloud data is of great importance in the cloud storage service. Considering the potentially huge number of on-demand data users and a lot of outsourced data documents in the cloud, this problem is particularly challenging as it is extremely difficult to meet also the requirements of performance, system usability, and scalability.

So, the data search on cloud storage should include smart indexing, de-duplication, and fast search with ranking. Ranked search can also elegantly eliminate unnecessary network traffic by sending back only the most relevant data from the storage. 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

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

completely protected. Besides, 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. Along 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 management and de-duplication techniques are reviewed.

A. Record Duplicate Detection Techniques:

Multiple versions of the same record are often accumulated when databases are constructed from multiple sources. The task of detecting these different versions is known as record de-duplication. Generally, the similarity of duplicate records is higher than the random pairs of records. All records that contain exactly or approximately the same data in one or more fields are identified in the process of duplicate detection. The problem of identifying syntactically different records that describe unique entities is denoted by all terms such as record linkage, duplicate detection and more. There are several approaches for solving duplicate detection problem [4]. Some of the approaches are Probabilistic Matching Models. This model uses a Bayesian approach to classify record pairs into two classes such as Supervised Learning and Unsupervised Learning. The supervised learning depends on the presence of training data in the form of record pairs, pre-labeled as same or not. In paper [5] authors used a popular CART algorithm generates classification and regression trees. A linear discriminate algorithm is also used to generate linear combination of the parameters for separating the data according to their classes and a vector quantization method which is a generalization of nearest neighbor algorithms. The idea behind the Unsupervised Learning technique is learning data to avoid manual labeling of the comparison vectors through clustering methods.

Active Learning based Techniques are introduced to detect the similar records, unlike an ordinary learner that trained using a static training data [6]. This technique uses active learner which actively selects subsets of instances from unlabeled data which, when labeled deliver the highest information gain to the learner. Distance Based learning, supervised learning and active learning techniques are not appropriate in the absence of training data. And these techniques are not fully optimized for the cloud storage data. One way of avoiding the essential of training data is to define a distance metric for records which does not need the training data. In the literature, the above stated approach detects the similar records is matched by using the distance metric and an appropriate matching threshold.

A number of researchers belonging to different groups that includes databases and machine learning have been studying this problem. The duplicate records under a single representative record are detected and then grouped or clustered in the duplicate record detection. De-duplication can be performed for a group of databases or for a single database that contains duplicate records. An ‘error free’ approach in the data warehouse is known as data quality. It is essential to enrich the quality of data through data cleaning methods.

Numerous data cleaning techniques are being employed for diverse 8 purposes. Similarities among records and fields are identified using Similarity Functions [7]. ‘Duplicate elimination functions’ are employed to identify if two or more records signify the same real world objects. Data cleaning methodologies which are in existence have been employed to recognize the missing values, record and field similarities and duplicate elimination [8]. As it is not possible to assume a unifying set of standards for various data sources, these issues are unavoidable. With increase in the size of the database, the problem intensifies. This is due to the huge amount of computational resource required for the examination and removal of duplicate records [9].

Duplicates can occur in numerous situations, for instance when a large database is updated by an external source and registry numbers are not accessible or are in error. Organizations are often confronted with the need to identify duplicate records present in huge databases. In a population register, there is a chance of some individual entities being listed under two or more registry numbers. Some information such as, name, address and date-of-birth may be necessary to identify the duplicates. Address or date-of-birth information is needed additionally as names do not uniquely identify. The identification of duplicates is difficult when names, addresses, and dates-of birth contain typographical errors [10]. Ahead of mining the accurate models, the data from the relevant sources must be collected,

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

integrated, cleaned and pre-processed in different ways. The merging of data from multiple databases into a single relation can often result in several duplicate records. These records are not syntactically identical, but, the same real-world entity is represented by them. To produce data of sufficient quality for mining, it is important to appropriately merge these records and the information they represent. Names such as record linkage, de-duplication, merge/purge, object identification, identity uncertainty, hardening soft information sources and more are also employed to denote this problem. The task of quickly and accurately identifying the records corresponding to the same entity from one or more data sources is known as record linkage [11]. The term approximate or fuzzy duplicates refer to the tuples which are somehow different but describe the same real world entity. The elimination of fuzzy duplicates in any database is necessary and yet it is vital in the data integration and analytical processing domains that require accurate reports/statistics [12]. The authors have utilized an approach for de-duplication that makes use of a fuzzy logic framework. The fuzzy inference system was optimized with the aid of the Bayesian Optimization Algorithm, a class of Estimation of Distribution Algorithms that are capable of learning complex multivariate relations of bounded order. Breeder genetic algorithms, utilized in the science of livestock breeding, were the motive behind the proposed class of algorithms. In Digital Libraries due to diverse sources of books that are spread across various parts of the country, duplicates could arise between scanning points. The Duplication of the books can be identified using only metadata of a book. If the metadata is missing, incorrect, abbreviated or incomplete it makes the duplicate detection all the more difficult. In paper [13] discussed a technique that works fast and efficiently in detecting the duplication of the books. Duplicate detection was done by similarity search using signature file method where the duplicates with typographical mistakes, word disorder, inconsistent abbreviations and even with missing words can be detected.

The performance of the similarity search is efficient since all the signatures are in the binary format and computations are done by low level logical operations. In paper [14] authors introduced a primitive operator SSJoin for performing similarity joins. They showed that similarity joins based on a variety of textual and non-textual similarity functions can be efficiently implemented using the SSJoin operator and then developed very efficient physical implementations for this operator mostly using standard SQL operators. 30 An adaptive framework to adaptively and dynamically regulate parameters of record linkage algorithms during the execution time was proposed in [15] Given a large collection of sparse vector data in a high dimensional space, in paper [16] investigated the problem of finding all the pairs of vectors whose similarity score (as determined by a function such as cosine distance) is above a given threshold. They proposed a simple algorithm based on novel indexing and optimization strategies that solve this problem without relying on approximation methods or extensive parameter tuning. It also showed that the approach efficiently handles a variety of datasets spanning a wide setting of similarity thresholds, with large speedups over previous state-of-the-art approaches. Several work presented a semantic text similarity measure aimed to guarantee uniqueness when used as an approximate join operator in the duplicate detection of a data cleaning process. By building sense matrices based on linguistic features, a semantic representation of the text is achieved. This representation is used as the input of the computation of the semantic similarity, which is expected to approximate texts with the same sense and to separate texts with distinct senses.

In [17], authors proposed a multi-level prefixfilter, to reduce the number of similarity calculations more efficiently and maintain the advantage of prefix-filter (no detection loss, no extra parameter) by applying multiple different prefix-filters. Several performance issues for a 31 software application that performs identification of duplicated records in a customer information database were presented in [18]. The utilization of edit distance measures the construction of a canonical representation that is central in the sense that it was most similar to each of the disparate records presented in [19]. The impact of noisy records on the canonical representation is reduced by this approach. They also learnt ranking preferences over canonicalization by exploring the feature-based methods. The approaches can select a canonical record by including arbitrary textual evidence in them. The real-world publications database was employed to evaluate the approach. The results illustrated that the learning method resulted in the canonicalization solution that was robust to errors and can be tailored in accordance with the user preferences with ease. In paper [20], authors proposed and compared two novel schemes for near duplicate image and video-shot detection. The first technique uses Locality Sensitive Hashing for fast retrieval and is based on global hierarchical colour histograms. The second one represents the image by a sparse set of visual words. The similarity is computed using a set overlap measure and efficient retrieval is achieved using a min-Hash algorithm. An approach to the technique called dynamic delayed duplicate detection has

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

been presented. Various typical properties of state spaces are exploited by the presented technique that also dynamically decides when to conduct duplicate detection by adapting itself to the structure of the state space. The experiments performed in an algorithm illustrates that the cost of duplicate detection was significantly reduced.

Authors in [21] proceeded towards duplicate elimination from a new perspective by considering de-duplication procedures as data processing tasks with tentative outcomes. They put forth an entirely uncertain model capable of efficiently encoding the space of clean instances of the input data, and instantiating efficient model implementations. The model was extended to capture the behavior of the de-duplication process, and support the revision and updation of the modeled uncertainty. The intricacy and scalability of the diverse techniques in different configurations was illustrated by the experimental evaluation.

The first step comprised of the training of examples of high quality that were automatically selected from the compared record pairs, and employing the same in the second step to train a Support Vector Machine (SVM) classifier. In [22], two variations of the approach were proposed. The first one worked on the basis of the nearest neighbor classifier whereas the second enhanced a SVM classifier by iteratively accumulating additional examples into the training sets. Authors in [23] have proposed a method that reports on experiments performed to investigate the use of a combined part of speech (POS) and an improved Longest Common Subsequence (LCS) in the analysis and calculation of similarity between texts. Later a method using hybrid mutation PSO algorithm is proposed to examine the optimized partial contents most similar in two documents.

B. Record Duplicate Detection over Unencrypted Data

Authors in [24] demonstrated the data de-duplication technique in terms of its negative issues, authors then proposed a randomized threshold method to avoid the negative impact. The authors used serer side duplication detection. But, the method proposed in the paper failed to utilize client-side data control proofs to prevent hash manipulation attacks in the storage services.

Later, based on the server-side de-duplication methods, the authors in [25] also demonstrated such negative issues on cloud storage. This technique uses de-duplication across multiple users and here, specifically it protects the hash values in the cloud storage. Because using the hash value the attacker may get all files. So, the authors protect the hash values by using PoW (proof of ownership). Using the PoW, a user should prove the server to get the values. Merkle trees are used in this paper rather than the short hash value.

On the basis of earlier study, authors in paper [25] also introduced a similar attack scenario on cloud storage that uses de-duplication across multiple users. Specifically, when an attacker temporarily compromises a server and obtains the hash values for data in the cloud storage, he is able to download all these data. This is because only a small piece of information about the data, namely, its hash value, serves as not only an index of the data to locate information of the data among a huge number of files, but also a “proof” that anyone who knows the hash value owns the corresponding data. Therefore, any users who can obtain the short hash value for specific data are able to access all the data stored in the cloud storage.

In paper [26] authors demonstrated the hash manipulation attack as like the earlier paper. The authors performed a practical evaluation of such attack in Dropbox. The Dropbox is one of the biggest cloud storage providers, so it practically proved the negative issues. However, the Dropbox folder provides full access to the malicious user to get the files of others if they know the index key.

Even though the existing de-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.

C. Record Duplicate Detection 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.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

Recently, authors in [27] 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 [28].

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 [30]. 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 [31] 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.

II. PROBLEM STATEMENT

One challenge in the cloud is finding that the relationship between documents will be normally concealed in the process of encryption, which will lead to significant search accuracy performance degradation. Also the volume of data in data centers has experienced a dramatic growth. This will make it even more challenging to design cipher-text search schemes that can provide efficient and reliable online information retrieval on large volume of encrypted data.

- The existing techniques is failed to perform document security in terms of integrity and dynamics analysis
- The indexing process was not completely studied for data search as well as data de-duplication.
- The searching accuracy in fully encrypted dynamic data is low
- Computational overheads are high

However, applying the privacy and security 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 index privacy, the keyword privacy, and many others.

Paper citation Number	Technique	Advantages	disadvantages
4	Probabilistic Matching Models	Data duplication detection is fast	Need to give proper training data
5	CART (classification and regression Tree) algorithm	Feature based data similarity detection. This is nonparametric, so does not rely on data belonging to a particular type of distribution.	Inability to handle multiple attributes
6	Active Learning based Techniques	Use static training data.	Couldn't detect for new record
7,8	Similarity Functions	Fast	it is not possible to assume a unifying set of standards for various data sources
12	Bayesian Optimization Algorithm	Optimized for data integration and analytical process	Require accurate test samples
13	signature file method	Used to find duplicate entries in the book domain	Not suitable for cloud

Table 1.0 comparative study of the data de-duplication techniques in data mining.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

De-duplication techniques can be categorized into two different approaches: de-duplication over unencrypted data and de-duplication over encrypted data. In the former approach, most of the existing schemes have been proposed in order to perform a de-dulicaiton process in an efficient and robust manner, since the hash of the file, which is treated as a "proof" for the entire file, is vulnerable to being leaked to outside adversaries because of its relatively small size. 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 distribution after the PoW process. However, not much work has yet been done to address dynamic ownership management and its related security problem.

III. 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 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. 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.

The research uses standard optimization algorithms such as GA, PSO and ABC for generating 36 the optimal similarity measure that decides whether the data is the same or not. Once the optimal similarity measure is obtained from the training data, the duplicate detection in testing dataset is carried out using the same measure. Improvements are made by including the exploration and exploitation process on the traditional PSO and ABC to provide even better performance and accuracy than the traditional one.

Finally, the work addresses the problem of 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.

IV. CONCLUSION

In this paper, the problem of finding and eliminating 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 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 and similarity matching techniques are 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 duplication 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. 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.
3. Hur, Junbeom, et al. "Secure data deduplication with dynamic ownership management in cloud storage." *IEEE Transactions on Knowledge and Data Engineering* 28.11 (2016): 3113-3125.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

4. Elmagarmid, Ahmed K., Panagiotis G. Ipeirotis, and Vassilios S. Verykios. "Duplicate record detection: A survey." *IEEE Transactions on knowledge and data engineering* 19.1 (2007): 1-16.
5. Cochinwala, Munir; Verghese Kurien and Gail Lalk and Dennis Shasha (2001). "Efficient data reconciliation". Information Sciences 137 (1-4): 1-15.
6. Sudhakaran, Saniya, and Meera Treesa Mathews. "A Survey on Data De-duplication in Large Scale Data." *International Journal of Computer Applications* 165.1 (2017).
7. Chiang, Yueh-Hsuan, AnHai Doan, and Jeffrey F. Naughton. "Modeling entity evolution for temporal record matching." *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*. ACM, 2014.
8. Randall, Sean M., et al. "The effect of data cleaning on record linkage quality." *BMC medical informatics and decision making* 13.1 (2013): 64.
9. Hamid HaidarianShahri, Saied HaidarianShahri, "Eliminating Duplicates in information Integration: An Adaptive, Extensible Framework", IEEE Computer Society 1541-1672, pp. 63-71, September/October 2006.
10. Winkler, William E. "Record linkage software and methods for merging administrative lists." *US Bureau of the Census* (2001).
11. Gu, Lifang, et al. "Record linkage: Current practice and future directions." *CSIRO Mathematical and Information Sciences Technical Report* 3 (2003): 83.
12. Shahri, Hamid Haidarian, and Ahmad Abdollahzadeh Barforush. "A flexible fuzzy expert system for fuzzy duplicate elimination in data cleaning." *International Conference on Database and Expert Systems Applications*. Springer, Berlin, Heidelberg, 2004.
13. L. Padmasree, V. Ambati, J. Chandulal, and M. Rao. Signature Based Duplication Detection in Digital Libraries. Signature, 2006.
14. 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.
15. 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.
16. 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.
17. Tateishi, Kenji. "Dai Kusui. Fast Duplicate Document Detection using Multi-level Prefix-filter." *The Third International Joint Conference on Natural Language Processing*. 2008.
18. 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.
19. 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.
20. Chum, Ondrej, James Philbin, and Andrew Zisserman. "Near Duplicate Image Detection: min-Hash and tf-idf Weighting." *BMVC*. Vol. 810. 2008.
21. Beskales, George, Mohamed A. Soliman, and Ihab F. Ilyas. *Modeling Uncertainty in Duplicate Elimination*. Technical Report, March 31, 2008.
22. 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.
23. 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.
24. 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.
25. 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.
26. 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.
27. 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.
28. J. Li, Y. K. Li, X. Chen, P. Lee, and W. Lou, "A hybrid cloud approach for secure authorized deduplication," IEEE Transactions on Parallel and Distributed Systems, Vol. 26, No. 5, pp. 1206–1216, 2015.
29. X. Jin, L. Wei, M. Yu, N. Yu and J. Sun, "Anonymous deduplication of encrypted data with proof of ownership in cloud storage," Proc. IEEE Conf. Communications in China (ICCC), pp.224-229, 2013.
30. M. Bellare, S. Keelveedhi, T. Ristenpart, "DupLESS: Serveraided encryption for deduplicated storage," Proc. USENIX Security Symposium, 2013.
31. Y. Shin and K. Kim, "Equality predicate encryption for secure data deduplication," Proc. Conference on Information Security and Cryptology (CISC-W), pp. 64–70, 2012.