



International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirccce.com

Vol. 5, Issue 3, March 2017

Duplicate Detection by Progressive Techniques

Kavya D., Rohini T.

P.G. Student, Dept. of CSE., New Horizon College Of Engineering, Bengaluru, India

Assistant Professor, Dept. of CSE., New Horizon College Of Engineering, Bengaluru, India

ABSTRACT: Data duplication is one of important data compression techniques for eliminating duplicate copies of repeating data, and has been widely used in cloud storage to reduce the amount of storage space and save bandwidth. To protect the confidentiality of sensitive data while supporting duplication, the convergent encryption technique has been proposed to encrypt the data before outsourcing. To better protect data security, this paper makes the first attempt to formally address the problem of authorized dataduplication. Different from traditional duplication systems, the differential privileges of users are further considered in duplicate check besides the data itself. We also present several new duplication constructions supporting authorized duplicate check. Security analysis demonstrates that our scheme is secure in terms of the definitions specified in the proposed security model. As a proof of concept, we implement a prototype of our proposed authorized duplicate check scheme and conduct testbed experiments using our prototype. We show that our proposed authorized duplicate check scheme incurs minimal overhead compared to normal operations.

KEYWORDS: Data Duplicity Detection; Entity resolution; Data cleaning;

I. INTRODUCTION

Duplication has been a well-known technique and has attracted more and more attention recently. Data duplication is a specialized data compression technique for eliminating duplicate copies of repeating data in storage. The technique is used to improve storage utilization and can also be applied to network data transfers to reduce the number of bytes that must be sent. Instead of keeping multiple data copies with the same content, duplication eliminates redundant data by keeping only one physical copy and referring other redundant data to that copy. Duplication can take place at either the file level or the block level. For file level duplication, it eliminates duplicate copies of the same file. Duplication can also take place at the block level, which eliminates duplicate blocks of data that occur in non-identical files.

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. Multiple, yet different of the same real-world objects in data, duplicates, are one of the most intriguing data quality problems. Several representations generally are not same and have certain differences like misspelling, missing values, changed addresses, etc. which makes the detection of duplicates very difficult. The detection of duplicates is very costly because the comparison among all possible duplicate pairs is required. For example in particular online retailers, offer huge catalogues comprising a constantly growing set of items from many different suppliers. As independent persons change the product portfolio, duplicates arise. While there is an obvious need for duplication, online shops without downtime cannot give traditional duplication.

Data has to be in integrity, if it exceeds the criteria, it is a duplicate. But due to data changes and sloppy data entry, errors such as duplicate entries might occur, making data cleaning and in particular duplicate detection indispensable. A user has little knowledge about the given data but still needs to configure the cleansing process. When user has only limited, maybe unknown time for data cleansing and wants to make best possible use of it. Then, simply start the algorithm and terminate it when needed. The result size will be maximized.

II. RELATED WORK

There are many researches which are carried out on the duplicate detection [1],[2] also known as entity resolution. But the most prominent algorithms are the progressive blocking[3] and then the progressive Sorted Neighbourhood Method [4].



International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirccce.com

Vol. 5, Issue 3, March 2017

The problem of merging multiple databases of information about common entities are frequently encountered in KDD [6] and decision support applications in large commercial and government organizations. The problem we study is often called the Merge/Purge problem and is difficult to solve both in scale and accuracy. Large repositories of data typically have numerous duplicate information entries about the same entities that are difficult to cull together without an intelligent “equation theory” that identifies equivalent items by a complex, domain-dependent matching process. We have developed a system for accomplishing this Data Cleansing task and demonstrate its use for cleansing lists of names of potential customers in a direct marketing-type application.

We explore a pay-as-you-go approach to entity resolution,[7] where we obtain partial results “gradually” as we perform resolution, so we can at least get some results faster. As we will see, the partial results may not identify all the records that correspond to the same real-world entity. Our goal will be to obtain as much of the overall result as possible, as quickly as possible. Entity resolution (ER) is the problem of identifying which records in a database refer to the same entity. In practice, many applications need to resolve large data sets efficiently, but do not require the ER result to be exact. For an example, people data from the web may simply be too large to completely resolve with a reasonable amount of work. As another example, real-time applications may not be able to tolerate any ER processing that takes longer than a certain amount of time. This paper investigates how we can maximize the progress of ER with a limited amount of work using “hints,” which give information on records that are likely to refer to the same real-world entity

[8]Efficient duplicate detection is an important task especially in large datasets. In this paper, they have compared two important approaches, blocking and windowing, for reducing the number of comparisons. Additionally, we have introduced Sorted Blocks which is generalization of blocking and windowing. Experiments with several real-world datasets show that Sorted Blocks outperforms the two other approaches. A challenge for Sorted Blocks is finding the right configuration settings, as it has more parameters than the other two approaches. An advantage of Sorted Blocks in comparison to the Sorted Neighbourhood Method is the variable partition size instead of a fixed size window. This allows more comparisons if several records have similar values, but requires fewer comparisons if only a few records are similar. In the future, one of our research topics will be to evaluate strategies that group records with a high chance of being duplicates in the same partitions.

Thorsten Papenbrock, Arvid Heise, and Felix Naumann[5] The limited records are found in same partition. By doing this the overall number of comparisons is reduced. The multi-pass method and transitive closure are used in blocking method. In windowing method, there are three phase. The first phase is to assign a sorting key to each record. Next phase is to sort the record based on key value. The final phase is to assume fixed window size and compare all pairs of records appear in the window. The multi-pass method performs the sorting and windowing approaches multiple times to avoid mis-sort due to error in the attributes. One of the advantages of using sorted block in comparing with sorted neighbourhood method is the variable partition instead of a fixed size window.

III. PROPOSED SYSTEM

To better protect data security, this paper makes the first attempt to formally address the problem of authorized data duplication. Different from traditional duplication systems, the differential privileges of users are further considered in duplicate check besides the data itself. We also present several new duplication constructions supporting authorized duplicate check in a architecture. Security analysis demonstrates that our scheme is secure in terms of the definitions specified in the proposed security model. As a proof of concept, we implement a prototype of our proposed authorized duplicate check scheme and conduct testbed experiments using our prototype. We show that our proposed authorized duplicate check scheme incurs minimal overhead compared to normal operations.

Advantages:

1. One critical challenge of storage services is the management of the ever-increasing volume of data. So by this approach we can find the solution which we are facing.

IV. IMPLEMENTATION

Implementation is the stage of the project when the theoretical design is turned out into a working system. Thus it can be considered to be the most critical stage in achieving a successful new system and in giving the user, confidence that the new system will work and be effective.

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

The implementation stage involves careful planning, investigation of the existing system and its constraints on implementation, designing of methods to achieve changeover and evaluation of changeover methods.

Main Modules:-

A. User Module:

In this module, Users are having authentication and security to access the detail which is presented in the ontology system. Before accessing or searching the details user should have the account in that otherwise they should register first. Figure 1 shows the screenshot of the user module UI. Here the user can insert the file which they want to store as in figure 2 and even they can view the file. Here the user can even view the duplicate records by sending request to the admin so that admin should accept to that request. Records can be even shared to the admin as seen the figure3. So after sharing if there is any duplicate record found it will throw an error message. Once the user is done with all his operations then they can sign out of that page.



Figure 1: The screenshot of user UI

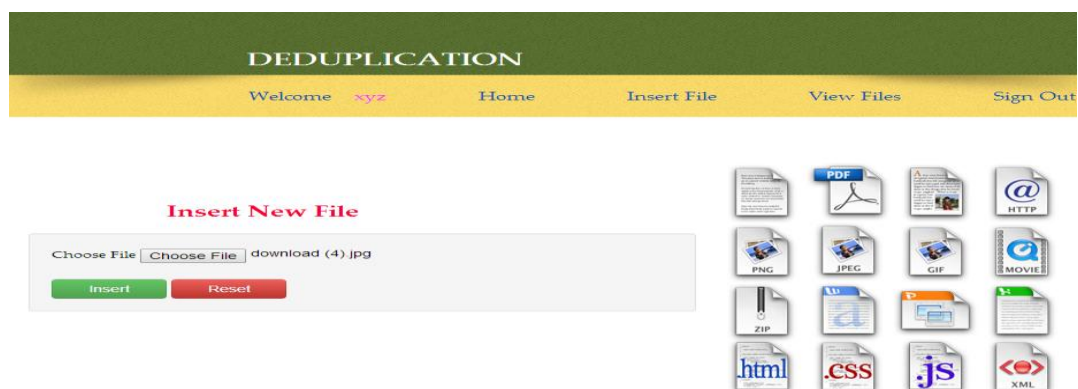


Figure 2: screenshot to show how the record insertion is done

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirccce.com

Vol. 5, Issue 3, March 2017

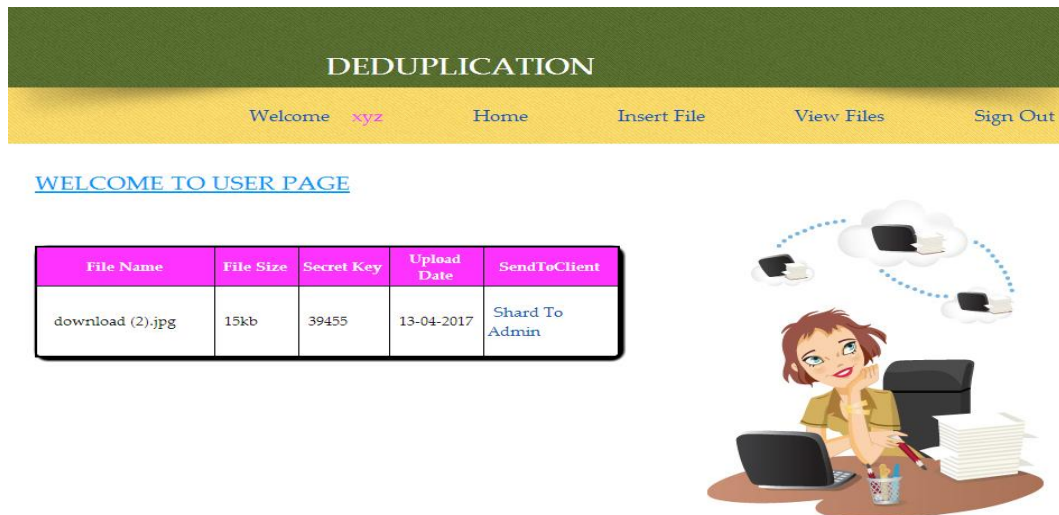


Figure 3:screenshot to show how can a user share record to admin

B. Admin module

This module contains user details where the admin can view the details of the user if the user is no longer active than he can detail that user from the database. The request whatever the admin got to view duplicate can be activated so that the user can view the duplicate record. Then the admin has a opportunity to view the list of records that are present in the database with all the shared details. The detail description will be know from the screen shot 4.

C. Secure Duplication System

To support authorized deduplication, the tag of a file F will be determined by the file F and the privilege. To show the difference with traditional notation of tag, we call it file token instead. To support authorized access, a secret key kp will be bounded with a privilege p to generate a file token. Let $\phi' F;p = \text{TagGen}(F, kp)$ denote the token of F that is only allowed to access by user with privilege p . In another word, the token $\phi' F;p$ could only be computed by the users with privilege p . As a result, if a file has been uploaded by a user with a duplicate token $\phi' F;p$, then a duplicate check sent from another user will be successful if and only if he also has the file F and privilege p . Such a token generation function could be easily implemented as $H(F, kp)$, where $H(_)$ denotes a cryptographic hash function.



Figure 4: UI of admin module

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

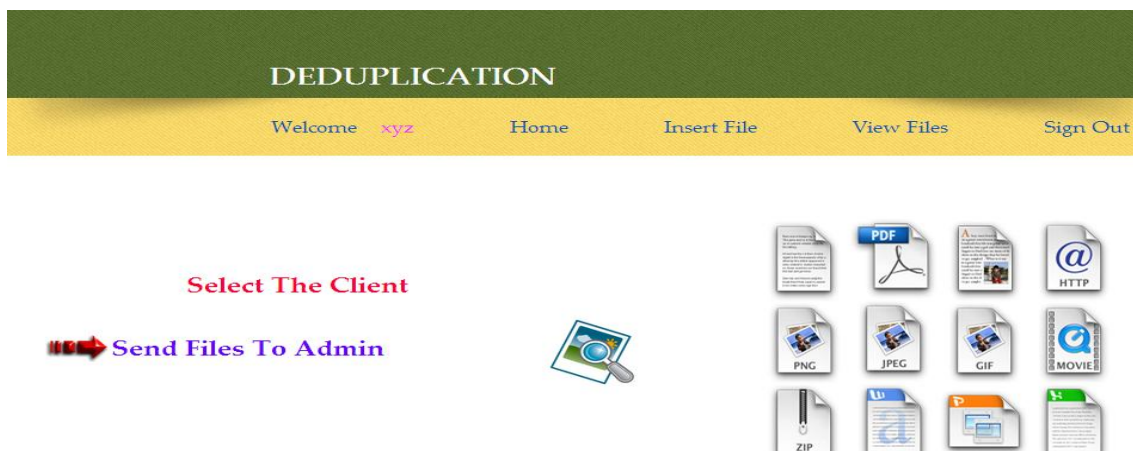
V. EXPERIMENTAL RESULTS

A quality output is one, which meets the requirements of the end user and presents the information clearly as in screen shot 5,6&7. In screenshot5 the user is sending record twice to admin so there will be a duplication as in shot6 then in next screen shot8 the admin will activate the request to view the record. Then in next shot8 the user can view the duplicate record. In any system results of processing are communicated to the users and to other systems through outputs. In output design it is determined how the information is to be displayed for immediate need and also the hard copy output. It is the most important and direct source of information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

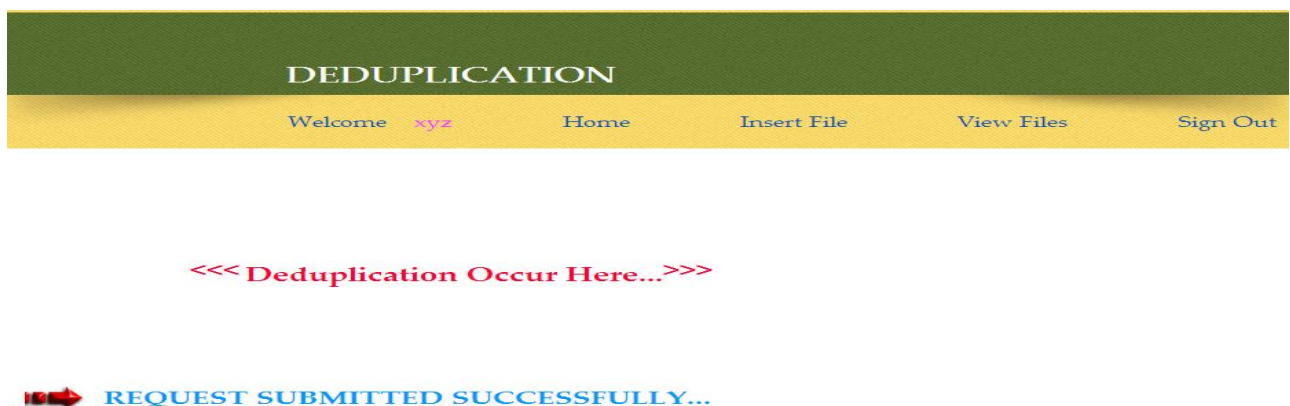
- 1) Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can be used easily and effectively. When analyzing design computer output, they should identify the specific output that is needed to meet the requirements.
- 2) Select methods for presenting information.
- 3) Create document, report, or other formats that contain information produced by the system.

The output form of an information system should accomplish one or more of the following objectives.

- Convey information about past activities, current status or projections of the future.
- Signal important events, opportunities, problems, or warnings.
- Trigger an action.
- Confirm an action.



Screenshot 5: sending record to admin for second time



Screen shot 6: shows that duplication occurred and requesting admin to view the record

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

DEDUPLICATION

Welcome [Admin](#) [Home](#) [Sign Out](#)

[ADMIN HOME](#)


User-ID	FileName	Status
kavya	download (1).jpg	REQUESTED
kavya	download (1).jpg	REQUESTED
xyz	download (2).jpg	REQUESTED
kavya	download (1).jpg	REQUESTED
xyz	download (2).jpg	REQUESTED
kavya	download (4).jpg	REQUESTED
kavya	download (4).jpg	REQUESTED
kavya	download (4).jpg	REQUESTED
kavya	download (4).jpg	REQUESTED
kavya	th.jpg	ACTIVATED

Screen shot 7: Activating request to view record

DUPLICATION

Welcome [xyz](#) [Home](#) [Insert File](#) [View Files](#) [Sign Out](#)

[THIS IS YOU REQUESTED DUPLICATES](#)

Screen shot 8: Duplicate record that user wants to view

VI. CONCLUSION AND FUTURE WORK

In this project, the notion of authorized data duplication was proposed to protect the data security by including differential privileges of users in the duplicate check. We also presented several new duplication constructions supporting authorized, in which the duplicate-check tokens of files are generated with private keys. Security analysis demonstrates that our schemes are secure in terms of insider and outsider attacks specified in the proposed security model. As a proof of concept, we implemented a prototype of our proposed authorized duplicate check scheme and conduct test bed experiments on our prototype. We showed that our authorized duplicate check scheme incurs minimal overhead compared to convergent encryption and network transfer.



International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirce.com

Vol. 5, Issue 3, March 2017

REFERENCES

- [1] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios, "Duplicate record detection: A survey," IEEE Transactions on Knowledge and Data Engineering (TKDE), vol. 19, page no. 1, 2007
- [2] F. Naumann and M. Herschel, An Introduction to Duplicate Detection. Morgan & Claypool, 2010.
- [3] H. B. Newcombe and J. M. Kennedy, "Record linkage: making maximum use of the discriminating power of identifying information," Communications of the ACM, vol. 5, page no. 11, 1962.
- [4] M. A. Hernández and S. J. Stolfo, "Real-world data is dirty: Data cleansing and the merge/purge problem," Data Mining and Knowledge Discovery, vol. 2, page no. 1, 1998.
- [5] Thorsten Papenbrock, Arvid Heise, and Felix Naumann "Progressive Duplicate Detection" IEEE Transactions on Knowledge and Data Engineering DOI 10.1109/TKDE.2014.2359666.
- [6] M. A. Hernández and S. J. Stolfo, "Real-world data is dirty: Data cleansing and the merge/purge problem, Data Mining Knowl. Discovery, vol. 2, no. 1, pp. 9–37, 1998.
- [7] U. Draisbach and F. Naumann, "A generalization of blocking and windowing algorithms for duplicate detection," in Proc. Int. Conf. Data Knowl. Eng., pp. 18–24, 2011
- [8] Steven Euijong Whang "Pay-As-You-Go Entity Resolution" IEEE transactions on knowledge and data engineering, vol. 25, page no. 5, may 2011

BIOGRAPHY

Ms Kavya D. Pursuing M Tech. Computer Science and Engineering in New Horizon College of Engineering, which is located in Outer Ring Road, Panathur Post, Kadubisanahalli, Bangalore – 560087.

Ms Rohini T. Assistant professor, Dept. of Computer Science and Engineering in New Horizon College of Engineering, which is located in Outer Ring Road, Panathur Post, Kadubisanahalli, Bangalore – 560087.