ISSN: 0975-766X
CODEN: IJPTFI
Available Online through *Research Article*
www.ijptonline.com

AN EFFECTIVE ANALYSIS AND SYSTEM BASED SAMPLING MODEL FOR DEDUPLICATION

K.Swathi Tejaswi*[1], A.K.Reshmy[2]
Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha University, Chennai.
Email:konakallaswathitejaswi@gmail.com

Received on 10-08-2016 *Accepted on 06-09-2016*

Abstract:

Data Deduplication is a method for eliminating duplicate copies of data, and has been widely used in data mining to avoid the repetition of data. The information provided by the user is to alter the duplication process usually represented by a set of physically labeled pairs. Two-stage Sampling Selection strategy (T3S), which is used for Reducing the set of pairs to avoid the deduplication Process in large Datasets using sample selection strategy and Redundancy removal. The T3S is mainly used to reduce the labeling effort while achieving the competitive quality when compared with matching and non-matching data. The training set is used to identify where the most ambiguous pairs lie and to configure the classification approach. Signature-based Deduplication is used efficiently to handle large deduplication tasks. The prefix filtering and length filtering is applied to remove records whose length variation is higher than specified. The Sampling Selection Strategy and Redundancy Removal Stages are used to avoid Deduplication. The report analysis is generated for the inputs. Certainly, this will be led by the ability to deduplicate unstructured data (office files, images, secured data etc.).

Keywords: Deduplication, Two-stage sampling strategy, Redundancy removal.

Introduction

We are going to show that how the large set of data can be processed along with the concept involved in deduplication using the T3S (Two-stage sampling selection)[12]. Data reduplication is the process of identifying references in data records that refer to the same real-world entity. Collective deduplication is a generalization in which one wants to find types of real-world entities in a set of records that are related. More corporate and private users outsource their data to cloud storage providers, recent data violate incidents make end-to end encryption an increasingly prominent requirement.

Unfortunately, semantically safe encryption schemes render various cost-effective storage optimization techniques, such as data deduplication, ineffective. It presents a novel idea that differentiates data according to their popularity. Deduplication is used to improve storage utilization and can also be applied to network data transfers to decrease the number of bytes that must be sent. In the deduplication process, unique chunks of data, or byte patterns, are recognized and stored during a process of analysis. Storage-based data deduplication reduces the amount of storage needed for a given set of files. It is most useful in applications where many copies.

Related Work

A. Two stage sampling strategy:

In this we integrate the T3S with the previous FS-Dedup framework to reduce the user effort in the main duplication steps[12][13]. Firstly we produce the candidate pairs. Secondly we remove the redundant data by means of rule-based active sampling [3][4]. These two steps work together to detect the fuzzy region [13]. Finally we describe the classification approaches.

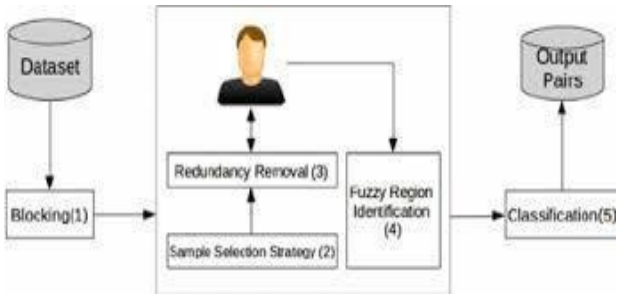

Fig.1. T3S steps overview.

B. Signature based deduplication: Preventing duplicate or near duplicate documents from entering an index or labeling documents with a signature/fingerprint for duplicate field collapsing can be efficiently achieved with a low collision or fuzzy hash algorithm. It can be achieved by using MD5 signature based algorithm[17][18]. When a document is added, a signature will automatically be generated and attached to the document in the specified signature Field[18].

Existing System: Baseline approach is inefficient, as it will generate an huge number of keys with the increasing number of users. The Existing system is unreliable. A semantic security for not accepted data and provides weaker security and improved storage and bandwidth for popular data. In this approach, data deduplication is the process by which a storage provider only stores a single copy of a file owned by several of its users. It only focuses an analysis on scenarios where the outsourced dataset contains few instances of some data items and many instances of others.

A. Problems in Existing System

Deduplication method aims at reducing the number of comparisons by grouping together pairs that share common features but not the unmatched pairs. In this proposed system, it is used to send only the random pairs of data and takes samples of each and every data simultaneously. The process is used only to avoid the labeling effort, but not the relevant data what they are going to process with help of sampling strategy.

Proposed System: The Proposed system uniquely shows that how a procedural analysis is maintained when the user need some amount of data. The deduplication is used here to overcome the repeated data at the earliest stage. We maintained each and every set of records using the statistic reports hence it is easier to identify the exact point where the user stands. It occurs limited overhead on normal upload/download operations .Proposed system supports both file-level and block level deduplication. Data deduplication will continue to make into other areas of storage, including archive and primary storage.

Experiment

A. Algorithm

SSAR (Selective sampling using association rule) it is a rule-based active selective sampling algorithm used to dissimilar unlabeled pair by making a comparison with the current training set[4].The unlabeled pair of dissimilar features produces a projection which generates few rules and if the dissimilar pair is not already present in the training set, it is labeled by the user[3]. The algorithm is the most effective way to achieve data security and eliminates repeated data also. When more pairs are labeled by the user, the training set becomes more informative and only the most dissimilar pairs are selected by the SSAR.

B. Architecture

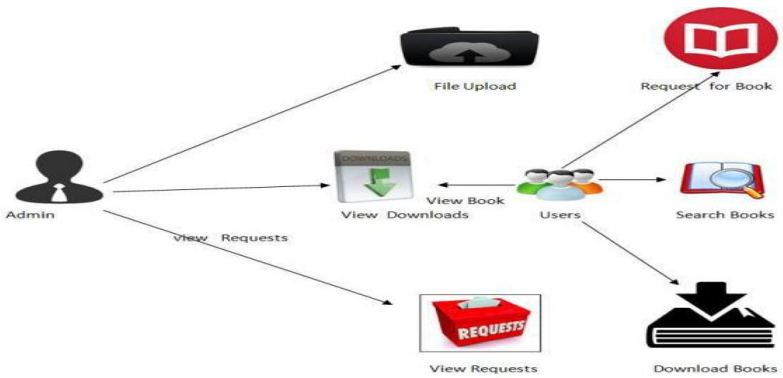

Fig.2. Selection process for deduplication.

In this we have taken an example of selecting the books from the library and performing deduplication process. We will see each step in a detailed process.

C. Catalogue Transfer

From Fig.1. File Uploading is the transmission of a file from one computer system to another, usually larger computer system. Here, the authorized users have the rights to upload a file, the repeated files with the same name or content cannot be uploaded, as the Deduplication occurs.

In order to reduce the space and size, the Deduplication concept is used. The repeated files will not be uploaded and those uploaded duplicate files will get eliminated and only the original data will be stored[1].If the same file is uploaded, it shows that the File name is already exists. If the same file is uploaded by renaming it shows that, it is a duplicate file.

D. Book Finder

The Book Finder is used to search the list of available books from a library so that the time can be reduced[14][15][16]. Here, we can search the Book by providing either of the text like Book Name, Author Name or Published Year[14]. The user needs approval from the administrator if user's downloads more than three Books and the approval will be sent to the particular user's mail and the user can download it again.

E. Visibility Exploiter

By eliminating physical handling and shelving of printed books as well as simplifying user searches, E-Books allow Administrator to reduce overhead and focus their efforts elsewhere. Administrator can view the number of Users Downloaded the File and their counts.

F. Users Invocation

User's invocation is to suggest the Admin to buy or provide the particular Book or a relevant Book to a library so that the Admin can able to know the necessary Books needed in a library and he/she will upload the same as early as possible. So the Registered Users can able to get the same and make use of it[17]. It eliminates damage, loss, and security concerns.

G. Statistics

The statistics displays the particular user's mail-id and the count of the book when that particular user is downloaded. Information is collected in a uniform way. They are usually easy to analyze. They are often required and respected by decision-makers. They overcome the difficulties of encouraging participation by users.

Future Work

Moving beyond backup, data deduplication will continue to make into other areas of storage including archive and primary storage. While the results will not be as exciting as the ones found in backup due to less data redundancy, the cost savings will still be considerable. Firstly, deduplication will help create another "tier" of primary storage between top-tier critical data and the backup tier. Undoubtedly, this will be led by the capability to deduplicate unstructured data (office files, images, secured data etc.)[9]. The more active nature of structured data (e.g., critical databases with strict performance requirements) means the road to deduplication usage for it will be longer-term.

Conclusion

In the existing system if two files have same name it gives an error as file already exists irrespective of the content present in it. If the file names are different and if the data inside those files are same, then it's a waste of space. This is duplication of data. In order to increase the space and reduce duplication the proposed system will match the data inside the files irrespective of file names. Even if the file names are same it does not matter. If the content is not similar it will save the data. The conclusion of the proposed system is to utilize space and increase the speed.

References

1. A. Arasu, M. Gotz and R. Kaushik , "On active learning of record matching packages", *Proc. ACM SIGMOD Int. Conf. Manage. Workshop QualityDatabases Manage. Data*, pp.783 -794 , 2010 *Uncertain Data* , pp.3 -12 , 2008.
2. A. Arasu, C. R¿¿ and D. Suciu , "Large-scale deduplication with constraints using dedupalog" , *Proc. IEEE Int. Conf. Data Eng.* , pp.952 -963 , 2009
3. R. J. Bayardo, Y. Ma and R. Srikant , "Scaling up all pairs similarity search" , *Proc. 16th Int. Conf. World Wide Web*, pp.131 -140 , 2007
4. K. Bellare, S. Iyengar, A. G. Parameswaran and V. Rastogi , "Active sampling for entity matching" , *Proc. 18th ACM SIGKDD Int.Conf. Knowl Discovery data mining. Pp. 1131-1139, 2012.*
5. M. Bilenko and R. J. Mooney, "On evaluation and training-set construction for duplicate detection" , *Proc. Workshop KDD* , pp.7 -12, 2003.

6. S. Chaudhuri, V. Ganti and R. Kaushik , "A primitive operator for similarity joins in data cleaning" , *Proc. 22nd Int. Conf. Data Eng.* , pp.5 , 2006.
7. P. Christen , "A survey of indexing techniques for scalable record linkage and deduplication" *IEEE Trans. Knowl. Data Eng.* , vol. 24 , no. 9 , pp.1537 -1555 , 2012
8. G. Dal Bianco, R. Galante, C. A. Heuser and M. A. Gonalves, "Tuning large scale deduplication with reduced effort", *Proc. 25th Int. Conf. Scientific Statist. Database Manage.* , pp.1 -12, 2013
9. A. Elmagarmid, P. Ipeirotis and V. Verykios , "Duplicate record detection: A survey" , *IEEE Trans. Knowl. Data Eng.* , vol. 19 , no. 1 , pp.1 -16 , 2007.
10. S. Sarawagi and A. Bhamidipaty , "Interactive deduplication using active learning" , *Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining* , pp.269 -278 , 2002
11. R. M. Silva, M. A. Gon¿¿alves and A. Veloso , "A two-stage active learning method for learning to rank" , *J. Assoc. Inform. Sci. Technol.* , vol. 65 , no. 1 , pp.109 -128 , 2014
12. J. Wang, G. Li and J. Fe , "Fast-join: An efficient method for fuzzy token matching based string similarity join" , *Proc. IEEE 27th Int. Conf. Data Eng.* , pp.458 -469 , 2011
13. L. Geng and H.J. Hamilton, "Interestingness Measures for Data Mining: A Survey," *ACM Computing Surveys,* vol. 38, no. 3, 2006.
14. H. Toivonen, M. Klemettinen, P. Ronkainen, K. Hatonen and H. Mannila, "Pruning and Grouping of Discovered Association Rules," *Proc. ECML Workshop Statistics, Machine Learning, and Knowledge Discovery in Databases,* pp. 47-52, 1995.
15. B. Lent, A.N. Swami and J. Widom, "Clustering Association Rules," *Proc. 13th Int',l Conf. Data Eng. (ICDE ',97),* W.A. Gray and P.-Å,. Larson, eds., pp. 220-231, 1997.
16. Y. Ishikawa, H. Kitagawa and N. Ohbo, "Evaluation of Signature Files as Set Access Facilities in OODBs," *Proc. ACM SIGMOD ',93,* P. Buneman and S. Jajodia, eds., pp. 247-256, 1993.
11. *Amphion. CS5315, High Performance Message Digest 5 Algorithm (MD5) Core. Datasheet.*