ISSN 2395-695X (Print)
ISSN 2395-695X (Online)
International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)
Vol. 2, Special Issue 13, April 2016

AN EFFECTIVE DIGITAL LIBRARY WITH DEDUPLICATION CONCEPT USING T3S STRATEGY

M. Sindhuja
B.E,Computer science and Engineering
Jeppiaar SRR Engineering College, Padur
Chennai, Tamilnadu
mrsindhuja95@gmail.com

S. Suganthi
B.E,Computer science and Engineering
Jeppiaar SRR Engineering College, Padur
Chennai, Tamilnadu
suganthiselvaraj12@gmail.com

T.R. Saravanan(Asst Professor)
B.E,Computer science and Engineering
Jeppiaar SRR Engineering College, Padur
Chennai, Tamilnadu
saravanan5_t_r@yahoo.co.in

***Abstract*— Data Deduplication is a technique for eliminating duplicate copies of data, and has been widely used in data mining to avoid the repetition of data's. 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 blocking and classification phases rely on the user to configure the process. The Training set is used to identify and to configure the classifications. The T3S is mainly used to reduce the labeling effort while achieving the competitive quality when compared with matching and non-matching data's. The Training set is used to identify where the most ambiguous pairs lie and to configure the classification approach. Signature-based Deduplication is 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*— *Data Deduplication,Signature based Dedup, T3S Strategy, Blocking.

I. INTRODUCTION

DATABASE plays an important role in today's Software field. Many IT companies,industries and systems depend on the accurate databases to carry out numerous operations[15]. Our analysis have ended up with a information that there is a dramatic growth in generation of data from different sources such as mobile devices and social media. Therefore, the information quality stored in the databases may have significant cost implications to a system. It relies on information to operate and conduct business. Working with an error-free data makes system memory more reliable.While merging databases ,there is a chance of storing redundant data in memory. However, the quality of data is degraded due to the presence of duplicate pairs with certain abbreviations,conflicting data, misspellings and redundant entities, among other problems[1].This should be avoided in the earlier stage in order to increase the efficiency of memory usage.

As the size of databases increases now-a-days, the methods of linkage and deduplication have become a tedious process to undergo[10].The quality of the data can be improved by detecting and avoiding the duplicate copies present in data repository using data deduplication technique[1]. Data Deduplication is a technique for eliminating duplicate copies of data in large datasets. It has been widely used in data mining to avoid the repetition of data.It has many real time applications like cloud paradigm, mail storage etc.,

Many techniques have been introduced under Data Deduplication concept. But all such techniques lack in their result as they constantly involves in string comparison. The strings are compared and they are considered as redundant data, if they match. This result produces a huge difference when a PDF File is compared[3]. The previous Deduplication techniques are inconvenient due to three major reasons, (1)Clustering algorithm clusters only single entity type which makes it difficult to answer multiple entities. (2)Users are not provided with valuable domain knowledge. (3)The adhoc way constraints makes the user impossible to answer their application needs.[16][17] [18][19][20][21].

In order to solve the issues generated by the existing system, we have proposed a T3S Strategy. Our proposed work is to apply effective Two Stage Sampling Selection (T3S) deduplication Strategy in Digital Library System in order to improve storage utilization. This strategy also checks the content in addition to the name of the file. So, the redundant copies can be eliminated at the earliest stage. In T3S Strategy, splits the file content into blocks.The two stages in T3S Strategy are,

1) Sample selection strategy.
2) Redundancy removal.

This strategy is used to examine the entire content of a PDF File to check whether it is a duplicate copy. First, the Sample Selection Strategy creates sample from the File. The

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)
International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)
Vol. 2, Special Issue 13, April 2016

PDF File is splitted into blocks which are represented as samples. Each and every block of the file to be checked is compared with the file that is already stored. Second, the Redundancy Removal stage compares both file and pairs the similar data. Dissimilar pairs are not taken into consideration. This exclusively saves our time by only concentrating on similar pairs.

II. METHODOLOGY

A *module* is a bounded contiguous group of statements having a single name and that can be treated as a unit. In other words, a single block in a pile of blocks.

- Catalogue Transfer
- Book Finder
- Visibility Exploiter
- Users Invocation
- Statistical Approach

Catalogue Transfer

- 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 size cannot be uploaded, as the Deduplication occurs.
- In order to reduce the space and size,the deduplication concept is used.
- The repeated files will be uploaded and those uploaded duplicate files will get eliminated and then only the original data's will be stored.

Book Finder

- The Book Finder is used to search the list of available Books from a library so that the time can be reduced.
- Here,we can search the Book by providing either of the values like Book Name,Author Name or published Year.

Visibility Exploiter

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

Users Invocation

- Users 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.
- It eliminates damage, loss, and security concerns.

Statistical Approach

Statistics are a useful method to investigate:

- Usage patterns
- Access patterns
- Resource provision
- Tracking trends and changes (by repetition over time)
- Performance of services

- Usage patterns here depicts the amount effective of utilization of application for Downloads and Request.This can be surveyed by the Admin for the regular maintanence.
- Each and every User once sign up will be provided with unique User name and Password for their access. Here User may refer the list of available books by searching under partical category. He/She can be request for file upload/download.
- The role of Admin is monitoring the User access.This happens while User downloads files ,giving request for the further downloads or User recommends for the unavailable books .This is maintained periodically in order to provide the statistical approach individual user.
- The service provided by the Admin role have to fulfill the User requirements. This is done by providing the regarding books in time, response immediately for the request, aware of Internet connectivity problem .And so performance measure has been depicted from both User and Admin side.
- Statistical approach provided mainly for limit checking.This can make the User as well as Admin to use the memory space in an efficient way.

III. SYSTEM ARCHITECTURE

Explanation

The new user enters the information in the registration page and a User ID with password is generated . The user enters the user name and password to login.Then, the Home Page is viewed. Here, the user can search the books in the book finder. The user can search the book by entering the book name, author name,category and published year. The books under the searched category are listed to the user view. The user can download the books he needed.

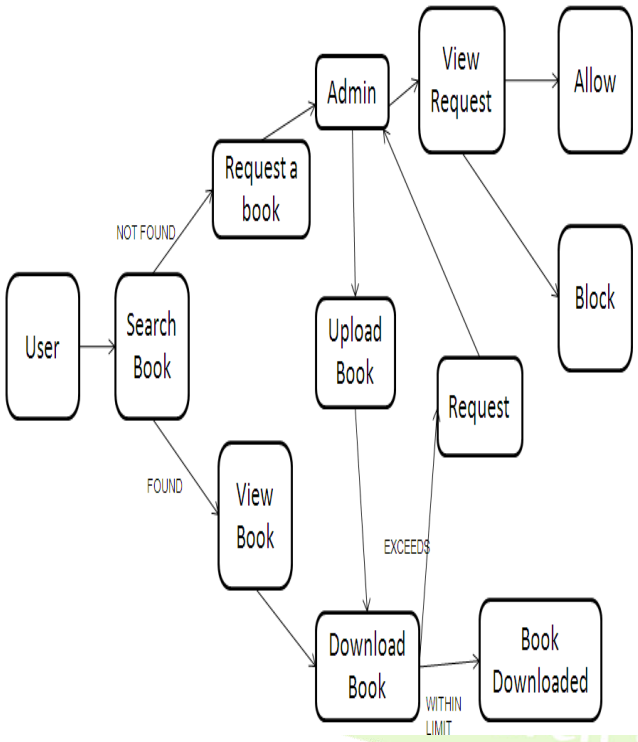
If certain books searched by the User are not found, then the books can be requested to the Admin. The Admin is also provided with a unique username and password. The Admin provides details like username and password to login. After login, he views the user request. The Admin after

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)
International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)
Vol. 2, Special Issue 13, April 2016

viewing the request, uploads the book to the website which is available to the User for download.

If the book with same name is uploaded/downloaded then a dialog box is opened which informs that the file name already exists. If the file name is renamed and proceded, then again T3S Strategy checks the content of the file. If the content matches, then again a dialog box emerges with a information that it is a duplicate copy. Here, the duplicate copies are identified and eliminated at the earliest stage to enrich the memory efficiency.

The Admin can set download limit to the users. The download limit is that the number of books ,the user can download within a particular time period. When the download limit of certain user exceeds, the information is sent to the Admin profile. Here, the Admin authenticates the user whether to allow or block the user for further downloads. The user can also request the admin to make downloads through feedback page.



IV. CONCLUSION

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 applies deduplication among convergent keys and distributes convergent key shares across multiple key servers and confidentiality of outsourced data. A new construction, T3S selects small random sub-samples of candidate pairs in different fractions of dataset.

V. FUTURE ENHANCEMENT

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 glamorous as the ones found in backup due to less data redundancy, the cost savings will still be substantial. Initially, deduplication will help create another "tier" of primary storage between top-tier critical data and the backup tier. Certainly, this will be led by the ability to deduplicate unstructured data (office files, images, secured data etc.). The more dynamic nature of structured data (e.g., critical databases with strict performance requirements) means the road to deduplication usage for it will be longer-term.

VI. REFERENCES

[1] C.A. Heuser, S.Canuto, R.Galante , " *A Practical and Effective Sampling Selection Strategy for Large Scale Deduplication*",in IEEE Trans on knowl and data engg, vol.27,pp.2305-2319, Sep. 2015.

[2] A. Arasu, M. Gotz, and R. Kaushik, "On active learning of record matching packages",in Proc. ACM SIGMOD Int. Conf. Manage. Data, pp.783–794, , 2010.

[3] A. Arasu, C. Re, and D. Suciu, "Large-scale deduplication with constraints using dedupalog," in Proc. IEEE Int. Conf. Data Eng., pp. 952–963,2009.

[4] R. J. Bayardo, Y. Ma, and R. Srikant, "Scaling up all pairs similarity search," in Proc. 16th Int. Conf. World Wide Web, pp. 131–140,2007.

[5] K. Bellare, S. Iyengar, A. G. Parameswaran, and V. Rastogi, "Active sampling for entity matching," in Proc. 18th ACM SIGKDD Int.Conf. Knowl. Discovery Data Mining, pp. 1131–1139, 2012.

[6] A. Beygelzimer, S. Dasgupta, and J. Langford, "Importance weighted active learning," in Proc. 26th Annu. Int. Conf. Mach.Learn., pp. 49–56, 2009.

[7] M. Bilenko and R. J. Mooney, "On evaluation and training-set construction for duplicate detection," in Proc. Workshop KDD, pp. 7–12, 2003.

[8] S. Chaudhuri, V. Ganti, and R. Kaushik, "A primitive operator for similarity joins in data cleaning," in Proc. 22nd Int. Conf. Data Eng.,p. 5, Apr. 2006.

[9] P. Christen, "Automatic record linkage using seeded nearest neighbour and support vector machine classification," in Proc.14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining,pp. 151–159, 2008.
[10] 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, Sep. 2012.
[11] P. Christen and T. Churches, "Febrl-freely extensible biomedical record linkage," Computer Science,Australian National University, Tech. Rep. TR-CS-02-05, 2002.
[12] D. Cohn, L. Atlas, and R. Ladner, "Improving generalization with active learning,"Mach. Learn., vol. 15, no. 2, pp. 201–221, 1994.
[13] G. Dal Bianco, R. Galante, C. A. Heuser, and M. A. Gonalves,"Tuning large scale deduplication with reduced effort," in Proc.25th Int. Conf. Scientific Statist. Database Manage., pp. 1–12, 2013.
[14] M. G. de Carvalho, A. H. Laender, M. A. Goncalves, and A. S. da Silva, "A genetic programming approach to record deduplication," IEEE Trans. Knowl. Data Eng., vol. 24, no. 3, pp. 399–412, Mar. 2012.
[15] A. Elmagarmid, P. Ipeirotis, and V. Verykios, "Duplicate record detection: A survey,"IEEE Trans. Knowl. Data Eng., vol. 19, no. 1,pp. 1–16, Jan. 2007.
[16] I. Fellegi and A. Sunter, "A theory for record linkage," J. Am. Stat-ist. Assoc., vol. 64, no. 328, pp. 1183–1210, 1969.
[17] S. Guha, R. Rastogi, and K. Shim, "Rock: A robust clustering algorithm for categorical attributes," Inf. Syst., vol. 25, no. 5, pp. 345–366, 2000.
[18] S. Chaudhuri, A. D. Sarma, V. Ganti, and R. Kaushik, "Leveraging aggregate constraints for deduplication," in SIGMOD Conference,pp. 437–448, 2007.
[19] W. W. Cohen, "Integration of heterogeneous databases without common domains using queries based on textual similarity." In SIGMOD Conference, pp. 201–212, 1998.
[20] A. K. H. Tung, R. T. Ng, L. V. S. Lakshmanan, and J. Han, "Constraint-based clustering in large databases," in ICDT, pp. 405–419, 2001.
[21] I. Bhattacharya and L. Getoor, "A latent dirichlet model for unsupervised entity resolution," in SDM, 2006.
[22] W. Shen, X. Li, and A. Doan, "Constraint-based entity matching," in AAAI, pp. 862–867, 2005.
[23] H. Pasula, B. Marthi, B. Milch, S. J. Russell, and I. Shpitser, "Identity uncertainty and citation matching," in NIPS, pp. 1401–1408, 2002.
[24] P. Andritsos, P. Tsaparas, R. J. Miller, and K. C. Sevcik, "Limbo:Scalable clustering of categorical data," in EDBT, pp. 123–146, 2004.
[25] I. Davidson, K. Wagstaff, and S. Basu, "Measuring constraint-set utility for partitional clustering algorithms," in PKDD, pp. 115–126, 2006.
[26] K. Wagstaff, S. Basu, and I. Davidson, "When is constrained clusteringbeneficial, and why?" in AAAI, 2006.
[27] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby, "Selective sampling using the query by committee algorithm,"Mach. Learn.,vol. 28, no. 2-3, pp. 133–168, 1997.
[28] C. L. Giles, K. D. Bollacker, and S. Lawrence, "Citeseer: An automatic citation indexing system," in Proc. 3rd ACM Conf. Digital Libraries, pp. 89–98, 1998.
[29] C. Gokhale, S. Das, A. Doan, J. F. Naughton, N. Rampalli, J. Shavlik, and X. Zhu, "Corleone: Hands-off crowd sourcing for entity matching," in Proc. ACM SIGMOD Int. Conf. Manage. Data,pp. 601–612, 2014.
[30] H. K €opcke and E. Rahm, "Training selection for tuning entity matching," in Proc. Int. Workshop Quality Databases Manage. Uncertain Data, pp. 3–12, 2008.
[31] S. Sarawagi and A. Bhamidipaty, "Interactive deduplication using active learning," in Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, pp. 269–278, 2002.
[32] R. M. Silva, M. A. Goncalves, 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.
[33] S. Tejada, C. A. Knoblock, and S. Minton, "Learning domain-inde-pendent string transformation weights for high accuracy object identification," in Proc. 8th ACM SIGKDD Int. Conf. Knowl.Discovery Data Mining, pp. 350–359, 2002.
[34] R. Vernica, M. J. Carey, and C. Li, "Efficient parallel set-similarity joins using mapreduce," in Proc. ACM SIGMOD Int. Conf. Manage.Data, pp. 495–506, 2010.
[35] J. Wang, G. Li, and J. Fe, "Fast-join: An efficient method for fuzzy token matching based string similarity join," in Proc. IEEE 27th Int. Conf. Data Eng., pp. 458–469, 2011.
[36] J. Wang, G. Li, J. X. Yu, and J. Feng, "Entity matching: How similar is similar," Proc. VLDB Endow., vol. 4, no. 10, pp. 622–633, Jul.2011.
[37] C. Xiao, W. Wang, X. Lin, J. X. Yu, and G. Wang, "Efficient similarity joins for near-duplicate detection,"ACM Trans. Database Syst.,vol. 36, no. 3, pp. 15:1–15:41, 2011.