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 2, February 2017

New Eminence Evaluation of Duplicate Detection

G. Sathish Babu[1], S.Sravani[2]

M.Tech Student, Dept. of CSE, St.Mark Educational Institution Society Group of Institutions, Affiliated to JNTUA, Andhra Pradesh, India[1]

Assistant Professor, Dept. of CSE, St.Mark Educational Institution Society Group of Institutions, Affiliated to JNTUA, Andhra Pradesh, India[2]

ABSTRACT: One of the serious problems faced in several applications with personal details management, customer affiliation management, data mining, etc is duplicate detection. This survey deals with the various duplicate record detection techniques in both small and large datasets. To detect the duplicity with less time of execution and also without disturbing the dataset quality, methods like Progressive Blocking and Progressive Neighborhood are used. Progressive sorted neighborhood method also called as PSNM is used in this model for finding or detecting the duplicate in a parallel approach. Progressive Blocking algorithm works on large datasets where finding duplication requires immense time. These algorithms are used to enhance duplicate detection system. The efficiency can be doubled over the conventional duplicate detection method using this algorithm. Several different methods of data analysis are studied here with various approaches for duplicate detection..

I. INTRODUCTION

Data mining is also called as KDD or knowledge discovery in database. The concept of data mining evolved from several researches that include statistics, database systems, machine learning concepts, neural networks, visualization, rough set, etc. Both traditional and latest areas like businesses, sports, etc use the data mining concepts. For translating the raw data into valuable information, the companies use a process. By knowing the details about the customers and by developing efficient marketing policies, the sales and costs can be increased or decreased in the businesses. The efficient collection of data, warehousing and computer processing all have their influence on data mining concepts. The data is the most essential important asset of any company but incase the data is changed or a bad data entry is made certain errors like duplicate detection arises. By using pattern recognition technologies and statistical and mathematical techniques to sift through warehoused information, data mining helps analysts recognize significant facts, relationships, trends, patterns, exceptions and anomalies that might otherwise go unnoticed. For businesses, data mining is used to discover patterns and relationships in the data in order to help make better business decisions. Data mining can help spot sales trends, develop smarter marketing campaigns, and accurately predict customer loyalty. Data mining technology can generate new business opportunities .

1.1.1Automated Prediction Of Trends And Behaviors:

Data mining automates the process of finding predictive information in a large database. Questions that traditionally required extensive hands-on analysis can now be directly answered from the data. A typical example of a predictive problem is targeted marketing. Data mining uses data on past promotional mailings to identify the targets most likely to maximize return on investment in future mailings. Other predictive problems include forecasting bankruptcy and other forms of default, and identifying segments of a population likely to respond similarly to given events.

DOI: 10.15680/IJIRCCE.2017. 0502098

1.1.2Automated discovery of previously unknown patterns:

Data mining tools sweep through databases and identify previously hidden patterns. An example of pattern discovery is the analysis of retail sales data to identify seemingly unrelated products that are often purchased together. Other pattern discovery problems include detecting fraudulent credit card transactions and identifying anomalous data that could represent data entry keying errors. While large-scale information technology has been evolving separate transaction and analytical systems, data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries. Several types of analytical software are available: statistical, machine learning, and neural networks. Generally, any of four types of relationships are sought:

Classes: Stored data is used to locate data in predetermined groups.
Clusters: Data items are grouped according to logical relationships or consumer preferences.
Associations: Data can be mined to identify associations.
Sequential patterns: Data is mined to anticipate behavior patterns and trends.

II. RELATED WORK

S. E. Whang et al. stated a survey on the active methods and non identical duplicate entries present in the records of the database records are all investigated in this paper. It works for both the duplicate record detection approaches. 1) Distance Based technique that measures the distance among the individual fields, by using distance metrics of all the fields and later computing the distance among the records. 2) Rule based technique that uses rules for defining that if two records are same or different. Rule based technique is measured using distance based methods in which the distances are 0 or 1. The techniques for duplicate record detection are very essential to improve the extracted data quality.

A.Thor et al. proposed a theory of deduplication which is also known as Entity Resolution which is used for determining entities associated to similar object of the real world. It is very important for data integration and data quality. Map Reduce is used for SN blocking execution. Both blocking methods and methods of parallel processing are used in the implementation of entity resolution of huge datasets.

U. Draisbach et al. in denoted a Duplicate Count Strategy is used which become accustomed to the window size depending on the count of duplicates detected. There are three strategies:

1. Key similarity strategy: The associations of the sorting keys influence the window size which is improved when the sorting keys are alike. Then we can expect several related records in this model.
2. Record similarity strategy: The associations of the records influence the window size. The replacement of the real resemblance of the records is present inside the window.
3. Duplicate count strategy: The count of the known duplicates influence the window size. DCS++ algorithm proves to be trustworthy than the SNM algorithm without losing the effectiveness. The algorithm of DCS++ is used to calculate the transitive closure and then save comparisons.

III. PROPOSED SYSTEM

The proposed system contains the existing system proposes also. In addition, the overall records are kept in multiple resources after splitting. The intermediate duplication results are intimated immediately after fount in any resources and are returned to the main application in proposed system. So the time consumption is reduced. Likewise the resource consumption is split across the resources.

A. ADVANTAGES

- The proposed system has following advantages.
- Concurrent approach is used. i.e., all the records are taken and checked as a parallel processes.
- Execution time is reduced.
- Resource consumption is same as existing system but the data is kept in multiple resource memories.

DESCRIPTION

A. DATASET COLLECTION

To collect and/or retrieve information concerning activities, results, context and alternative factors. it's vital to contemplate the sort of data it need to assemble from your participants and therefore the ways in which you may analyze that information. the information set corresponds to the contents of one info table, or one applied math information matrix, wherever each column of the table represents a specific variable. when aggregation the information to store the info**.**

B. PREPROCESSING METHOD

Data Preprocessing or information improvement, information is cleaned through processes like filling in missing values, smoothing the wheezy information, or resolution the inconsistencies within the information. And additionally accustomed removing the unwanted information. ordinarily used as a preliminary data processing follow, information preprocessing transforms the information into a format which will be a lot of simply and effectively processed for the aim of the user.

C. DATA SEPARATION

After finishing the preprocessing, the information separation to be performed. The block algorithms assign every record to a set cluster of comparable records (the blocks) and so compare all pairs of records inside these groups. every block inside the block comparison matrix represents the comparisons of all records in one block with all records in another block, the equal block, all blocks have constant size.

D. DUPLICATE DETECTION

The duplicate detection rules set by the administrator, the system alerts the user concerning potential duplicates once the user tries to form new records or update existing records. to keep up information quality, you'll be able to schedule a replica detection job to examine for duplicates for all records that match a particular criteria. you'll be able to clean the information by deleting, deactivating, or merging the duplicates rumored by a replica detection

E. QUALITY MEASURES

The quality of those systems is, hence, measured employing a cost-benefit calculation. particularly for ancient duplicate detection processes, it's troublesome to satisfy a budget limitation, as a result of their runtime is difficult to predict. By delivering as several duplicates as potential in an exceedingly given quantity of your time, progressive processes optimize the cost-benefit quantitative relation. In producing, a live of excellence or a state of being free from defects, deficiencies and vital variations. it's caused by strict and consistent commitment to sure standards that bring home the bacon uniformity of a product so as to satisfy specific client or user needs.

IV. CONCLUSION

This paper offered the progressive sorted local procedure and modern blockading. Each algorithms increase the efficiency of duplicate detection for instances with restrained execution time; they dynamically trade the ranking of evaluation candidates situated on intermediate outcome to execute promising comparisons first and not more promising comparisons later. To determine the performance gain of our algorithms, we proposed a novel quality measure for progressiveness that integrates seamlessly with present measures. For the development of a thoroughly revolutionary replica detection workflow, we proposed a modern sorting approach, Magpie, a innovative multi-go execution model, Attribute Concurrency, and an incremental transitive closure algorithm. The variations AC-PSNM and AC-PB use multiple type keys simultaneously to interleave their modern iterations. By analyzing intermediate outcome, both approaches dynamically rank the different variety keys at runtime, greatly easing the important thing resolution obstacle. In future work, we wish to mix our modern approaches with scalable approaches for duplicate detection to supply outcome even rapid. In specified, Kolb et al. Introduced a two section parallel SNM, which executes a traditional SNM on balanced, overlapping partitions. Here, we can rather use our PSNM to step by step in finding duplicates in parallel

REFERENCES

[1] S. E. Whang, D. Marmaros, and H. Garcia-Molina, “Pay-as-you-go entity resolution,” IEEE Trans. Knowl. Data Eng., vol. 25, no. 5,pp. 1111–1124, May 2012.

[2] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios, “Duplicate record detection: A survey,” IEEE Trans. Knowl. Data Eng., vol. 19,no. 1, pp. 1–16, Jan. 2007.

[3] F. Naumann and M. Herschel, An Introduction to Duplicate Detection.San Rafael, CA, USA: Morgan & Claypool, 2010.

[4] H. B. Newcombe and J. M. Kennedy, “Record linkage: Making maximum use of the discriminating power of identifying information,” Commun. ACM, vol. 5, no. 11, pp. 563–566, 1962.

[5] M. A. Hernandez 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.

[6] X. Dong, A. Halevy, and J. Madhavan, “Reference reconciliation in complex information spaces,” in Proc. Int. Conf. Manage. Data,2005, pp. 85–96.

[7] O. Hassanzadeh, F. Chiang, H. C. Lee, and R. J. Miller,“Framework for evaluating clustering algorithms in duplicate detection,” Proc. Very Large Databases Endowment, vol. 2, pp. 1282–1293, 2009.

[8] O. Hassanzadeh and R. J. Miller, “Creating probabilistic databases from duplicated data,” VLDB J., vol. 18, no. 5, pp. 1141–1166, 2009.

[9] U. Draisbach, F. Naumann, S. Szott, and O. Wonneberg,“Adaptive windows for duplicate detection,” in Proc. IEEE 28th Int. Conf. Data Eng., 2012, pp. 1073–1083.

[10] S. Yan, D. Lee, M.-Y. Kan, and L. C. Giles, “Adaptive sorted neighborhood methods for efficient record linkage,” in Proc. 7th ACM/IEEE Joint Int. Conf. Digit. Libraries, 2007, pp. 185–194.

[11] J. Madhavan, S. R. Jeffery, S. Cohen, X. Dong, D. Ko, C. Yu, and A. Halevy, “Web-scale data integration: You can only afford to pay as you go,” in Proc. Conf. Innovative Data Syst.Res., 2007.

[12] S. R. Jeffery, M. J. Franklin, and A. Y. Halevy, “Pay-as-you-go user feedback for dataspace systems,” in Proc. Int. Conf. Manage. Data,2008, pp. 847–860.

[13] C. Xiao, W. Wang, X. Lin, and H. Shang, “Top-k set similarity joins,” in Proc. IEEE Int. Conf. Data Eng., 2009, pp. 916–927.

[14] P. Indyk, “A small approximately min-wise independent family of hash functions,” in Proc. 10th Annu. ACM-SIAM Symp. Discrete Algorithms, 1999, pp. 454–456.Fig. 10. Duplicates found in the plista-dataset.1328 IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 27, NO. 5, MAY 2015

[15] U. Draisbach and F. Naumann, “A generalization of blocking and windowing algorithms for duplicate detection,” in Proc. Int. Conf.Data Knowl. Eng., 2011, pp. 18–24.