International Journal of Research
Available at https://edupediapublications.org/journals
p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 04 Issue 09
August 2017

Algorithms for Efficient Duplicate Detection

D.SaiSree & N.Sirisha
PG Student, Dept. of MCA, QIS College of Engineering &Technology , VengamukkalaPalem.
Assistant Prof, Dept. of MCA, QIS College of Engineering &Technology , VengamukkalaPalem.

ABSTRACT:

The presence of duplicate file is main data quality concern in large databases. To identification duplicates attribute resolution is known as duplication detection. Duplicate detection is process of detecting all cases of multiple method of same real world entity. In this paper propose new naive detection algorithm using intelligent guesses which records have a high possibility of representing the same real-world entity, the search space is reduced. An implement naive algorithm is used as base line identification generates all possible many objects is stored within the datasets. A new e-mail abstraction scheme is proposed to consider e-mail layout structure to display e-mails. New security a Robust and Collaborative Spam Detection System is used which possesses if efficient near duplicate matching scheme and a progressive update scheme. To maintain data quality and schedule a duplicate detection for all records is match a certain new process. Data clean deleting deactivating or merging the duplicates files by change detection method. The proposed system identification for detection of duplication data by using ranking methods.

Index Terms: Data cleaning, Record linkage, Naive Bayes, HTML, Pay-As-You-Go, data cleaning, dcs++.

1. INTRODUCTION

Today databases play an important role in IT based economy. Number of industries and systems depend on the efficiency of databases to take total operations. The quality of the records that are stored in the databases is significant cost indications to a system the relies on information to conduct business [1]. With this ever increasing bulk of data, the data quality problems Duplicate files detection is divided into three steps. Candidate description, to decide which objects is compared with each other [2] And secondly duplicate definition, the criteria based on which two duplicate candidates are in reality duplicates. The duplicate detection problem has two aspects: First, the multiple representations are usually not the same but contain differences, such as misspellings, changed addresses, or missing values. This makes it difficult to detect these duplicates. Second, duplicate detection is a very expensive operation, as it requires the comparison of every possible pair of duplicates using the typically

complex similarity calculate. The paper proposes the Parallel Duplicate Detection with Map reduce concept [3]. The adaptive techniques improve the efficiency in detecting the duplication but these techniques cannot bear up to the level of progressive techniques. The Progressive techniques could process larger dataset in short span of time and the quality of data is also good comparatively [5]. The Progressive duplicate detection makes it different from the traditional approach by yielding more complex results during the early termination [4] . The algorithms of duplicate detection also computes the duplicates at an almost constant frequency but the progressive algorithms increase the overall time as it finds out the duplicates at the early stage itself. The candidate keys in the record pairs that are identical have to be first found out. the next stage of spam detection research should focus on dealing with cunning spams which evolve naturally and continuously. In this paper, a novel e-mail abstraction scheme is proposed which considers e-mail layout structure to represent e-mails[6]. A procedure to generate the e-mail abstraction using HTML content in e-mail is presented, which can more effectively capture the near-duplicate phenomenon of spams. Moreover, a complete spam detection system is designed, which possesses an efficient near-duplicate matching scheme and a progressive update scheme. The progressive update scheme enables system to keep the most up-to-date information for near-duplicate detection [7].

2. RELATED WORK

Many research on duplicate detection [8] also named as entity resolution gives different methods for pair selection and duplicate detection of the records. One of the most important algorithms in this area are Blocking [9] and sorted neighborhood method (SNM).Blocking methods divides the data records into disjoint subsets, while windowing methods, in specific the Sorted Neighborhood Method, slide a window over the sorted records and compare records within each window. And we had an algorithm called Sorted Blocks [10] in several variants, which generalizes both the approaches. A challenge for Sorted Blocks is in finding the right configuration settings, as it has more parameters than the other two approaches. A merit of Sorted Blocks compared to the Sorted Neighborhood Method is the variable partition size instead of a fixed size window. This let more comparisons if different records have same values, but requires lesser comparisons if only a few records are similar. In Record matching [12] algorithms the existence of duplicate records constitutes a problem which is becoming increasingly alarming in networked environments, as the size of individual databases increases and new cooperative networks or consortia are created. Special algorithms are

developed for this purpose. The integrity of bibliographic databases are maintained with the algorithms of record matching. The similar records could never be matched at any stage despite matching the bibliographic descriptions. The progressive techniques like pay-as-you-go [11] algorithms were used for integration on large scale datasets. In pay-asyou-go, we theoretically order the candidate pairs by the chances of a match. The ER algorithms are used for performing this.Entity resolution (ER) is the problem of identifying which records in a database refer to the same entity. For real-time applications ER processing takes longer than a certain amount of time.The progress of ER can be maximised using hints that give information on records that are likely to refer to the same real-world entity. A hint can be represented in various formats and ER uses this information as a guideline for which records are computed first [13]

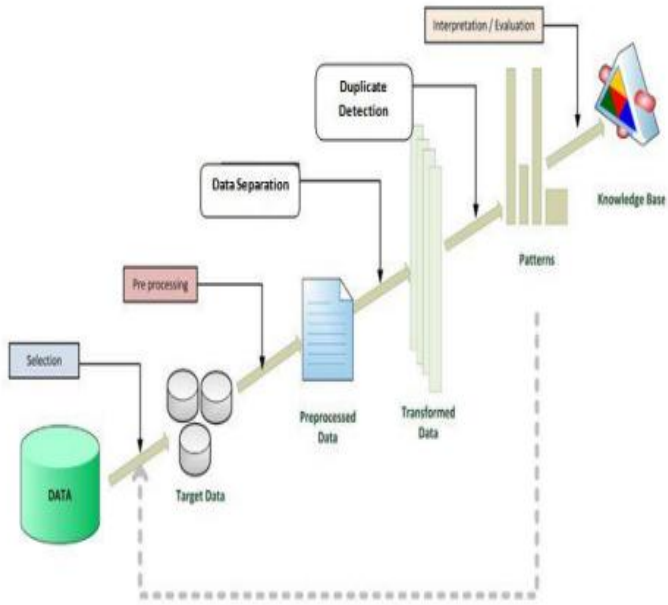

Figure1. Detection of Duplicate Datasets

3. PROGRESSIVE SORTED NEIGHBOURHOOD METHOD

The process of duplicate detection is the method of identifying multiple representations of same real world entities. Today, duplicate detection methods need to process very larger datasets in very shorter time: maintaining the quality of a dataset becomes increasingly difficult. One existing system for finding duplicates includes progressive duplicate detection method.

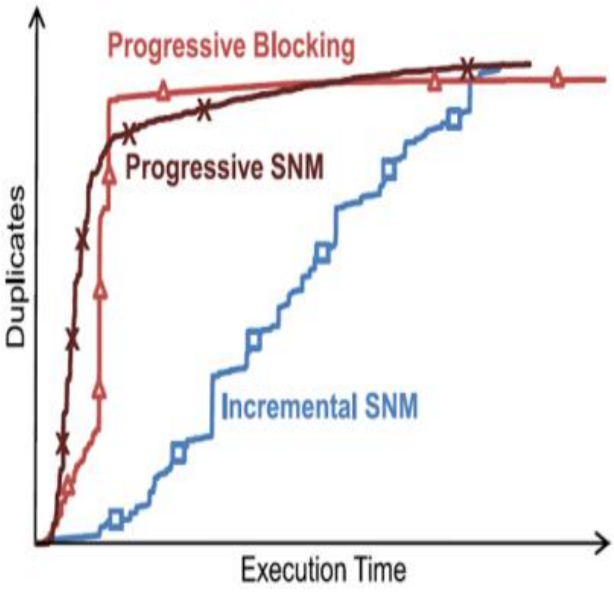

Fig 2: Duplicates the two progressive algorithm

a window. The perception is that data records that are close in the sorted order are more likely to be duplicates than records that are far apart, because they are already alike with respect to their sorting key.

The PSNM algorithm differs by dynamically changing the execution order of the comparisons based on look-ahead results. Progressive blocking (PB) algorithm [14] is another method for duplicate detection. It is a blocking algorithm instead of windowing method. Progressive blocking (PB) is an approach that initiates upon an equidistant blocking technique and the successive enlargement of blocks. Even though the progressive algorithms and the snm method give faster results it may not find accurate number of duplicates for large datasets. So this disadvantage can be solved by using the proposed algorithm.

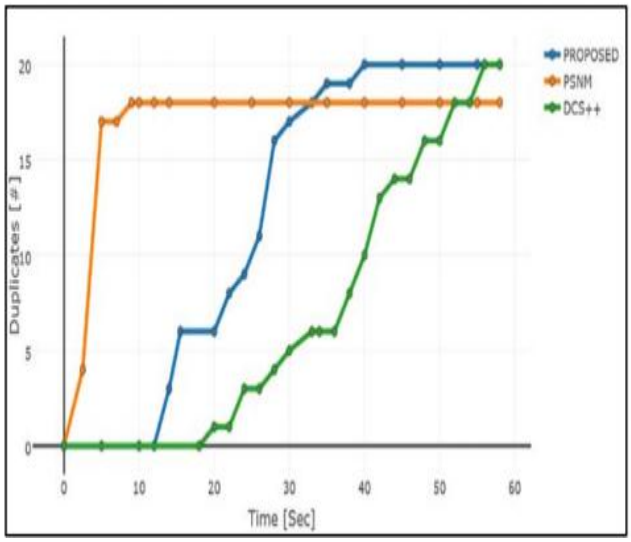

Fig 3: Comparison of duplicates found by psnm

4. PROPOSED SYSTEM

To overcome the problem of serial duplicate detection this work proposes an efficient and flexible detection scheme that supports both Progressive duplicate detection with map reduce and parallel duplicate detection. The proposed system is based on Map Reduce Algorithm.

A. **Training Dataset**:- In this Process user give the input data to the proposed system. Here training dataset loaded from company database or inserting from user

B. **Data Preprocessing**:-Data preprocessing is a data mining technique that involves transforming raw data into an understandable format[15]. Real-world data is often incomplete,

inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors

C. **Data Separation**: In this process we separate the large amount of data large data cannot be fit in to main memory so it is divided into different parts each part is called as cluster

D. **Duplicate Detection**: In this process we detect the duplicate records from cluster

The most important task is duplicate detection. The process of identifying multiple representation is generally named as duplicate detection. Generally, a particular user may rate the product two or more times. This multiple representation is named as duplication. The process of duplicate detection generally comprises three steps such as pair-selection, pair wise comparison and clustering. Here the efficiency of duplicate detection is improved by the concurrent approach. The detection and elimination of duplication is done concurrently [16].

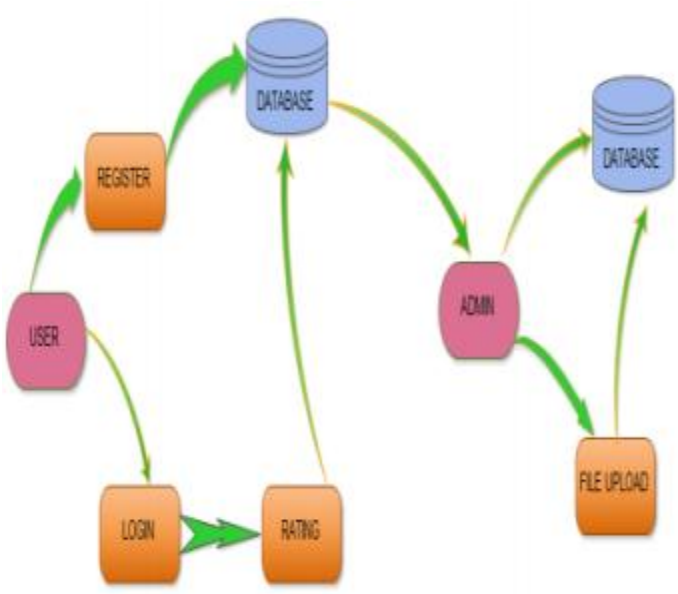

Figure 4.Detetion of duplicate datasets using algorithm

A. The System Model Codes

The system model of the algorithmic form is outlined starting to do the spam detection, Cosdes collects feedback spams for time Tm in advance to construct an initial database. Three major modules, Abstraction Generation Module, Database Maintenance Module, and Spam Detection Module, are included in Cosdes. With regard to Abstraction Generation Module, each e-mail is converted to an e-mail abstraction by Structure Abstraction Generator with procedure SAG. Three types of action handlers, Deletion Handler, Insertion Handler, and Error Report Handler, are involved in Database Maintenance Module. In addition, Matching Handler in Spam

Detection Module takes charge of determining results.

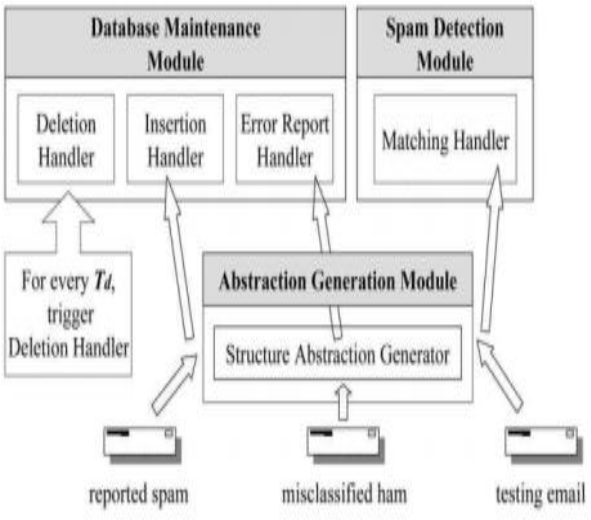

Fig. 5. System model of Cosdes.

There are three types of e-mails, reported spam, testing email, and misclassified ham, required to be dealt with by Cosdes. When receiving a reported spam, Insertion Handler adds the e-mail abstraction of this spam into the database except that the reputation score of this reporter is too low.

Spammers are finding ways to trick people into thinking their unsolicited junk messages are worth the time you spend reading them. A list of the top five ways to tell if an email is spam is as follows [17]. These rules can help you when spam slips through the protection of your Spam filter.

• If it ends up in Spam Folder

• Look at the Email Address

• Look at the Content

• If it asks for personnel Information

• Look at the Greeting

B. PROGRESSIVE METHOD WITH ADAPTIVE WINDOW

We propose a new method which is a combination of progressive sorted neighborhood method and data count strategy (dcs++). And this method helps to overcome some of the demerits of this algorithm. In this system window enlargement process is not used as in the progressive sorted neighborhood method instead uses sorting, partitioning and other methods but it uses the windowing and partitioning concept of dcs++. Here the main concept used is the partitioning and distance calculation thus finding the duplicates. The whole data record is partitioned into different partitions of same size and duplicate detection is done with the partitions. Thus it takes slighter more time than the progressive sorted neighborhood method but yields better results by detecting more number of duplicates. And when compared to dcs++, the processing speed of the proposed system is less thus overcoming the disadvantage. Even though compared to psnm it gives more duplicates and take less time than the dcs++ algorithm, the proposed method still takes time. So the processing speed can be increased

and this issue can be solved by using map reduce technique to this algorithm as it is done with snm to provide parallel sorted neighborhood method[18].

5. RESULT EVALUATION

Detection of duplication was very arduous in the beginning. Later on many algorithms were proposed to detect the duplication, but each algorithm had its‘ own drawbacks. At each stage, certain technologies were used with advancement at each stages. In this paper, the advancement brought was to detect the duplication concurrently along with several other operations. The graph indicates the improvement of performance in detecting the duplication over a period of time drastically.

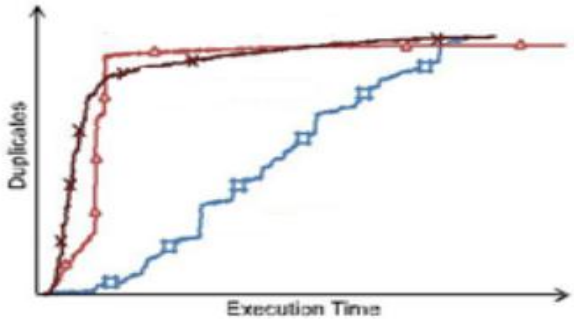

Figure 6: Performance evaluation for detection of duplication

6. CONCLUSION AND FUTURE WORK

This paper introduces an algorithm that is used for detecting duplication concurrently. This algorithm increase the efficiency of duplicate detection for situations with limited execution time; they dynamically change the ranking of comparison candidates based on intermediate results to execute promising comparisons first and less promising comparisons later. The specific procedure SAG is proposed to generate the e-mail abstraction using HTML content in email, and this newly-devised abstraction can more effectively capture the near-duplicate phenomenon of spasm. In future work, we want to connect our progressive duplicate detection using adaptive window algorithm with scalable methods for duplicate detection to provide results even faster. a complete spam detection system Cosdes has been designed to efficiently process the near-duplicate matching and to progressively update the known spam database.

7. 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] U. Draisbach, F. Naumann, S. Szott and O. Wonneberg, “Adaptive Windows for Duplicate Detection”, Proceedings of the IEEE 28th International Conference on Data Engineering, Arlington, Virginia, USA, (2012) April 1-5

[3] 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, no. 1, 1998

[4] Progressive Duplicate Detection‖, Thorsten Papenbrock, Arvid Heise, and Felix Naumann, IEEE Transactions on Knowledge and Data Engineering,May 2015.

[5] Efficient and Effective Duplicate Detection Evaluating Multiple Data using Genetic Algorithm, Dr.M.Mayilvaganan, M.Saipriyanka, Sep 2015

[6] M. Siva kumar reddy & Krishna Sagar. "Improved Near duplicate matching scheme for e-mail spam Detection" International Journal of Internet Computing ISSN No: 2231 – 6965, VOL- 1, ISS- 4 2012 29

[7] S. Sarafijanovic, S. Perez, and J.-Y.L. Boudec, "Resolving FP-TP Conflict in Digest-Based Collaborative Spam Detection by Use of Negative Selection Algorithm," Proc. Fifth Conf. Email and Anti-Spam (CEAS), 2008.

[8] 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.

[9] 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.

[10] 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.

[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]. R. Baxter, P. Christen, and T. Churches, ―A comparison of fast blocking methods for record linkage,‖ in Proceedings of the ACM SIGKDD Workshop on Data Cleaning, Record Linkage, and Object Consolidation, 2003, pp. 25–27.

[13]. S. E. Whang, D. Marmaros, and H. Garcia-Molina, ―Pay-as-you-go entity resolution,‖ IEEE Trans. Knowl. Data Eng., vol. 25, no. 5, May 2012.

[14] Thorsten Papenbrock, ArvidHeise, and Felix Naumann, "Progressive Duplicate Detection," IEEE Transactions on Knowledge and data engineering, vol. 27, no. 5, May 2015.

[15]. 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.

[16]. L. Gu and R. Baxter, ―Adaptive filtering for efficient record linkage,‖ in Proceedings of the SIAM International Conference on Data Mining, 2004, pp. 477–481.

[17] E. Damiani, S.D.C. di Vimercati, S. Paraboschi, and P. Samarati, "P2P-Based Collaborative Spam Detection and Filtering," Proc. Fourth IEEE Int'l Conf. Peer-to-Peer Computing, pp. 176-183, 2004.

[18] L. Kolb, A. Thor, and E. Rahm, "Parallel sorted neighborhood blocking with MapReduce," in Proc. Conf. Datenbanksysteme in B€uro, Technik und Wissenschaft, 2011.

AUTHORS:

D.SaiSree is currently pursuing her **MCA** in MCA Department, **QIS college of Engineering & Technology, VengamukkalaPalem** , A.P. She received her Bachelor of Computer Applications from **ANU.**

N.Sirisha was completed her MSC. Presently she is working as an Assistant Professor in MCA Department, **QIS College Of Engineering&Technology, VengamukkalaPalm.** Her research includes networking and data mining.