ISSN:0975-9646

A Survey on Various Duplicate Detection Methods

M.Priyanka[#1], Asha Baby[#2]
[1]*Final year M. Tech CSE, Vimal Jyothi Engineering college, kannur,Kerala.*
[2]*Asst.Prof.Department of CSE, Vimal Jyothi Engineering, college, Kannur,kerala.*

Abstract: **Data growth is tremendously increasing day by day,as the data growth increases there are more chances to have duplicates hence it is difficult to maintain the quality of data. Duplicate detection is the process of representing same real world Entities in multiple ways. it is also known as record relation, record linkage, Entity classification, record matching. Duplicate detection is a necessary task in data cleansing and relevant for data integration. Two well -known algorithms in this area are blocking and windowing. In this literature survey various techniques used to find duplicate records are described.**

Keywords: **Duplicate Detection,Data Mining,Data cleaning,Progressiveness.**

I. INTRODUCTION

Data are the important asset of all organization. Duplicate detection is serious problem in many applications, including customer relationship management, personal information management. Whenever we want to consider a duplicate detection from dataset which mostly comes under Data mining.Data mining is something like extracting information from a large tons of data.It has attracted a great deal of attention in the society and information industry and in society as a whole in recent years.

We are rich in terms of data but the information is poor until turning such data into use full information there is no use of such data so we need to mine the relevant information data growth is exponentially increasing so human ability cannot understand or extract information from data without using powerful tools. So there is a need of data mining .While dealing with data mining we suppose to know about knowledge discovery from data (KDD).It consisting of several steps such as data cleaning, integration, data selection,transformation,data mining pattern evaluation and finally knowledge presentation. First four steps are different forms of data preprocessing. Typically, the process of duplicate detection is followed by a data preparation stage during which data entries are stored in a uniform manner in the database and partially resolving the structural heterogeneity problem. With the increasing volume of data, data quality problems arise. Duplicates are one of the most interesting data quality problem.Effects of such duplicates are harmful, for example If one person is now at goa but his hometown is kerala.Then single person consider as two persons when data collected by this two different states. So data gets duplicated.Now this scenario considered for thousands of people in every state, that time very large number of data get duplicated so there is an obvious need for duplicate detection.

Duplicate detection is the process of identifying various representations of a same real-world objective in a information source. Nowadays duplicate detection methods need to process larger datasets in shorter time. maintaining the quality of a dataset becomes increasingly difficult.Duplicate detection problem has two aspects. First one is multiple representations that are usually not the same but contain differences, such as changed addresses, misspellings, or missing values. This makes difficulties to detect these duplicates. Second duplicate detection is a very expensive operation, as it requires the comparison of every possible pair of duplicates typically complex to calculate similarity of every pair of record.

Blocking and Windowing are the two approaches used in duplicate detection. Windowing is Sorted neighborhood method that compare sorted records within window when it slides. Blocking approach is based on partion record method. Progressiveness improves the results, efficiencies and scalability of the algorithms used in the existing model. progressive approaches try to reduce the average time after which a duplicate is found. Early termination, in particular, that gives more complete and accurate results.

Main aim of this survey paper is to research about several duplicate detection approaches. Brief summary about different methods are described in section II, method comparison are described in section III, at last conclusion of review paper is described in section IV.

II. LITERATURE REVIEW

Paper [8] Described sorted neighbourhood approach[SNM]. This Method consists of following three steps first one is Creating a key for each record in the list is computed by extracting relevant fields or portions of fields.Second is Sorted data.The records in the database are sorted by using the key found in the first step. sorting key is defined to be a sequence of attributes or a sequence of substrings within the attributes chosen from the record. Third step is Windowing where a Fixed size window slides over the sorted data. All pairs of records within a window are compared and duplicates are marked**.**

Entity resolution (ER) [6] is the problem of identifying records in a database refer to the same entity. Many applications need to resolve large data sets efficiently but do not require the ER result to be exact. Real-time applications may not be able to tolerate any ER processing that takes longer than a certain amount of time. Introduced technique called hints to maximize the progress of ER with a limited amount of work. Gives information

about records that are likely to refer to the same real-world entity. A hint can be represented in various formats .Three types of hints that are well-matched with different ER algorithms as sorted list of record pairs , a hierarchy of record partitions and an order list of records ER can use this information as a guideline for which records to compare first .

Distance and rule based approach is used in [3] Distance-Based Approach is used to calculate the distance between specific fields using the proper distance metric for each field. Later computing the distance among the records . Rule based Approach is a special case of distance-based approaches. It uses rules to define whether two records are the same or not. Rule based approach can be dignified as distance-based approach Where the distance between two records represented in single bit binary number.

Paper [4] Described about sorted blocks that provides a generalization of blocking and windowing methods Blocking methods split records into separate partitions and perform complete comparison within these partition. Sorted Neighbourhood Method use sliding window over the arranged records and compare records only within the window. Advantage of Sorted Blocks in comparison to the Sorted Neighbourhood Method is the variable partition size instead of a fixed size window.

In [2] Duplicate Count Strategy is used, which adapts the window size based on the number of detected duplicates

Characterizing the strategies by defining the movement of the right and left boundaries of the window. Window size increases by advancing the right boundary and it decreases by advancing left boundary. Adaptation can increase or reduce number of comparisons. Three strategies are used. Key similarity strategy in which window size is improved, if sorting keys are similar and thus more related records can be expected. Record similarity strategy, Window size is varied based on the similarity of the records. Duplicate count strategy where window size is based on the number of well-known duplicates. Adapts the window size for each and every duplicate in the current window.

Paper [5] described top-k similarity join that uses a special index structure to estimate promising comparison candidates. Comparing record pairs in the order of their similarity. Finding a pair of records such that their similarities are not less than a given threshold. List of duplicates almost ordered by similarity. Drastically reduce the candidate size.

Duplication detection is also called as Record Linkage Used a window to build non-overlapping blocks that can contain different numbers of records. The pairwise record comparison then takes place within these blocks. Hypothesis is that the distance between a record and its successors in the sort sequence is monotonically increasing in a small neighbourhood. Used two algorithms Incrementally Adaptive-SNM (IA-SNM) is an algorithm that increases the window size Incrementally. Accumulative Adaptive-SNM creates windows with one overlapping record. Both IA-SNM and AA-SNM, compared with SNM through experimental evaluation. Both approaches do not conform that it perform better than SNM [7].

In [1] Developed two dynamic progressive duplicate detection algorithms progressive sorted neighbourhood method (PSNM) and progressive blocking (PB). PSNM and PB dynamically adjust their behaviour by automatically choosing optimal parameters, like window sizes, block sizes, and sorting keys, rendering their manual specification .The progressive sorted neighbourhood method is based on the traditional sorted neighbourhood method. PSNM algorithm differs by dynamically changing the execution order of the comparisons based on intermediate results .Progressive blocking approach that builds upon an equidistant blocking technique and the successive enlargement of blocks.

III. METHOD COMPARISON

Method	Data Set Used	Problems Detected	Reference
SNM	Personal information	1)Due to the fixed window size, it cannot scale well. 2) it cannot efficiently respond to different size within a data set	[8]
Sorted Blocks	Restaurant Data	1) Difficult to find the right configuration settings 2) It has more parameters than the SNM and blocking	[4]
Adaptive versions of SNM IA-SNM AA-SNM	Cora Data Restaurant Data	1)Not fully adapted 2)Adaptation can increase or reduce number of comparisons	[7]
Duplicate Count Strategy	Personal data Cora Data Set	1)Skipping windows may miss duplicates 2)Difficult to calibrate	[2]
PSNM PB	CD Data	1)Failure in detection of duplicates in the edges of partitions	[1]
Hints	Restaurant Data	1)Static order of comparisons 2) calculate a hint only for a specific partition	[3]

VI. CONCLUSION

Various duplicate detection approaches are studied in this paper. The existing techniques which have algorithms to detect duplicity in records improve the competence in finding out the duplicates when the execution time is less. The process gain within the available time is maximized by reporting most of the results.

REFERENCES

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

[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 1-5C.

[3] 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.

[4] 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.

[5] Xiao, W. Wang, X. Lin, and H. Shang, "Top-k set similarityjoins," in Proc. IEEE Int. Conf. Data Eng., 2009, pp. 916–927

[6] 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.

[7] 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.

[8] M. A. Hernandez and S. J. Stolfo, "Real-world data is dirty: Data cleansing and the merge/purge problem," Data Mining Knowledge Discovery, vol. 2, no. 1, pp. 9–37, 1998.