iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 2 **Issue:** XI **Month of publication:** November 2014

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

A Review on Effective and Efficient Detection of Duplicates in Data

Varsha Wandhekar [#1], Arti Mohanpurkar [*2]
Department of Computer Engineering, Dr.D.Y.Patil School of Engineering & Technology,

Abstract— Detection of Duplication is an essential step in data cleansing. Data duplication techniques are used to link records which relate to the same entity in one or more data sets where a unique identifier is not available. Duplication detection is also called as Record Linkage. The major challenges in detection of duplication are the computational complexity and the linkage accuracy. Blocking and Windowing are two approaches used in Duplication detection. Windowing is a Sorted Neighbourhood Method; compare the records within window when it slides. Blocking is partition record method. The main focus of this paper is on maintain and improve efficiency as well as effectiveness of duplication detection by using adaptive windowing and blocking algorithms.
Keywords— Duplication Detection, Record linkage, Sorted Neighbourhood Method (SNM), Similarity Measure, Windowing, Blocking

I. INTRODUCTION

In an every organization or companies have present numerous sites, and each of them generates large amount of data. If we consider Election system then data is collected from all states. If one person is now at Maharashtra but his hometown is Bangalore, then single person consider as two persons when data collected by this two different states. So data get duplicated. Now this scenario considered for thousands of people in every state, that time very large number of data get duplicated. So finding out the same entity from different data sources is very important and difficult task.

Data duplication detection is the problem of identifying record pairs that represent same real world entity and could this merged into single record. Duplication detection is an important component of data cleansing and integration. Cleansing of data is one of the most decisive steps. If data is dirty, inaccurate, incomplete, and inconsistent then decisions taken on the basis of data may be misleading or not good. The existence of duplicates is major issue in dirty data. An important cleansing issue is removal of duplicates [1], [15].

There are two types of record matching; the first is structural heterogeneity and the second is lexical heterogeneity. The problem of matching two databases with different domain structures is Structural heterogeneity. For e.g. a customer address stored in the attribute 'address' in one database but represented in attributes 'street', 'city', and 'zip' in another database. The databases with similar structure but different representation of data are Lexical heterogeneity, such as 'V. Roy', 'Vijay R.' and 'Roy, Vijay' [2].

A number of data mining tasks involve computing similarity between pairs of records. The total number of pairwise similarity computations grows gradually with the size of the input dataset, scaling to large datasets is problematic task.

For small datasets, estimation of the full similarity matrix can be difficult. The most instance pairs are highly dissimilar so in many task majority of similarity computation are unnecessary. Blocking methods only selects a subset of record pairs for similarity computation, ignoring the remaining pairs which are irrelevant and highly dissimilar [3].

In the past years numbers of blocking algorithms have been proposed by researchers [3], [4], [5], [6], [7], [10]. These techniques typically form blocks or groups of observations using sorting or indexing. For subsequent similarity computations this allows efficient selection of instance pairs from each block. Some blocking methods are based on the similarity metric. The most common reason of mismatch in database entries is typographical variation of String data. To deal with typographical variations in duplication detection typically relies on String comparison techniques. Various methods have been developed for this task. Character-Based Similarity Metrics are: Edit distance, Jaro-Winkler distance, Smith-Waterman distance, Q-gram distance, Affine gap distance [8], [9], [15].

The most important representative for windowing is Sorted Neighborhood Method (SNM). It has three phases:

A. Key selection: Sorting key is assigned to each record. The key is generated by concatenating two or more values of attributes.
B. Sorting: All records are sorted according to key.
C. Windowing: Slides a window over sorted data. Within particular window all records pairs are compared and duplicates are marked.

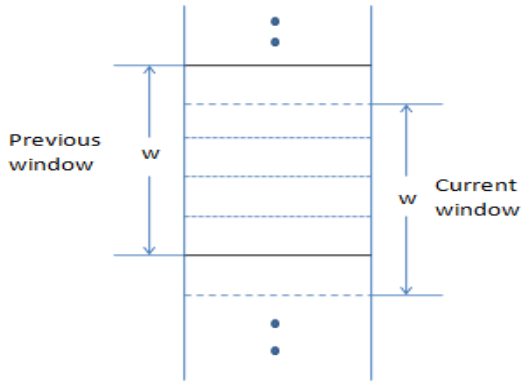


Fig.1: The representation of Window in Sorted Neighborhood Method

A disadvantage of the Sorted Neighborhood Method is the fixed window size. Some duplicates might be missed when selected window size is too small. On the other hand, unnecessary comparison carried out when window size too large. To achieve effectiveness adaptive window size is used [3], [7], [11], [12], [14].

As an election system entities that duplication detection problem must deal with more became large scaled and heterogeneous. In order to make duplication detection solution applicable, we consider that adaptively plays important role. So in this paper we focuses on adaptively and dynamically changing parameters of duplication detection during execution. To maintain effectiveness and efficiency we compare the Incrementally Adaptive SNM (IA-SNM), accumulatively adaptive SNM (AA-SNM) and sorted block algorithms [7],[10],[11].

II. RELATED WORK

Duplication detection is use for reduce cost and search space. Su Yan discuses record linkage using 'adaptivity' algorithms of SNM. They compare different number of records within blocks by using non-overlapping blocks. The sorting of records is done by lexicographically not by distance. They present IA-SNM and AA-SNM this two algorithms and compare them with basic SNM.

Incrementally Adaptive-SNM (IA-SNM) is an algorithm that increases the window size incrementally. Although calculate the distance of the first and the last element in the current window and if this distance is smaller than threshold then window get enlarger. Enlargement of window size is depends on the current window size.

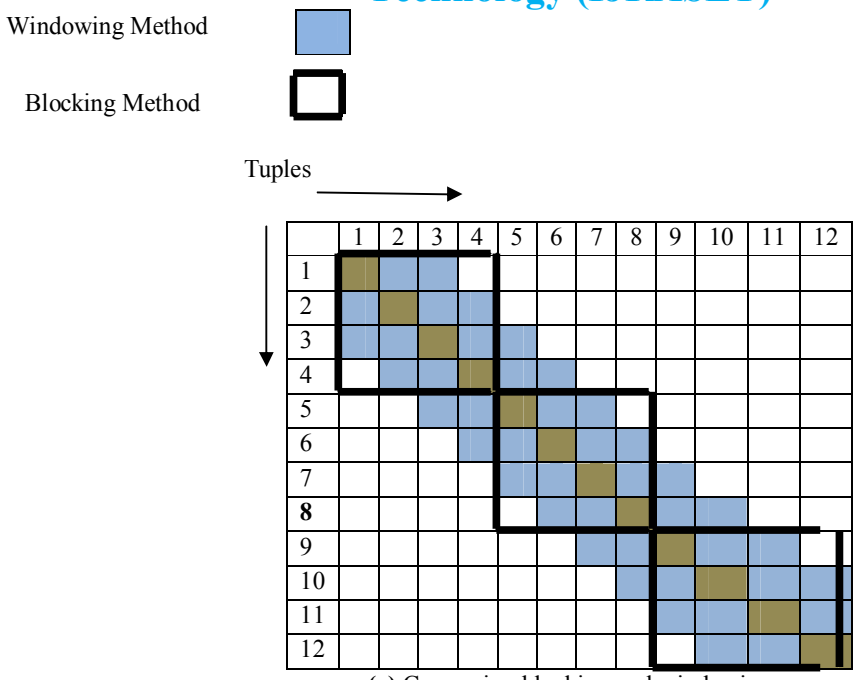
On the other hand accumulative Adaptive-SNM creates windows with one overlapping record. By consideration, if last record of a window is a potential duplicate of the last record in next adjacent window then multiple adjacent windows can be grouped into one block. Enlargement of window is depends on distance of the first and the last record comparing with threshold. Both algorithms have retrenchment phase after the enlargement phase. Retrenchment of window size is until all records within the block are potential duplicates.

Fig.2 shows the schematic representation of both sorting and windowing method under a common model. In this comparison of two tuples shows in cell of matrix. The only centre diagonal cells perform comparison [11].

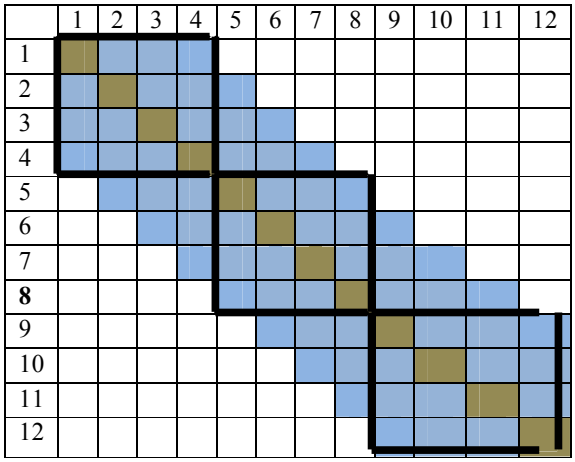
Figure 2(a) shows the blocking and windowing method. This method shows that same number of comparison when approximately calculated. But actual comparisons are different. For example, In blocking method tuple 2 and 4 are compared only, because they lie in the same partition. In windowing method only tuple 4 and 5 are compared.

Fig.2 (b) shows that comparison of windowing method when window size was increased.

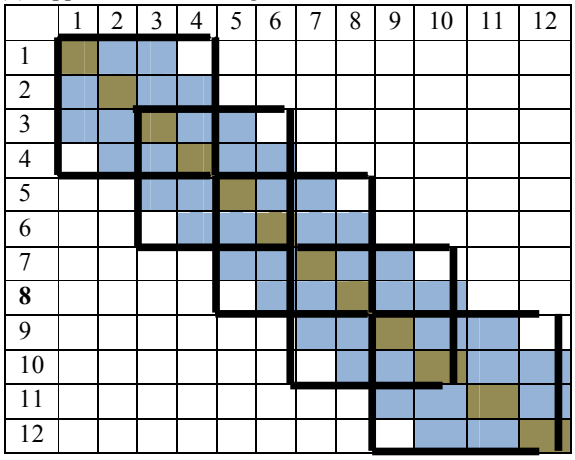
Fig.2(c) shows blocking method adapt by different partitions so all comparison made as per windowing method [7], [10], [12].



(a) Comparing blocking and windowing



(b) Approximate blocking when increase the window size



(c) Approximate windowing when overlap the blocks

Fig. 2: shows the schematic representation of both sorting and windowing method under a common model. Each field in matrix corresponds to a comparison of two tuples.

Uwe discuss Sorted Blocks variants. In this method all records are sort according to sorting key. After that disjoint partitions are creating and all records within that particular partition are comparing. In this additionally an overlapping partition is also used to ensure duplication in different partitions. The size of overlapping partition is assign as O_v, eg O_v=2 that means two tuples from each neighboring partition are overlap part. Within window, the records comparing within each window that are fixed window size is O_v + 1. Sometimes fixed size partition results give missed duplicates, so adaptive size partition is better approach according to partition predicate. [7]

III. GENERAL DUPLICATION DETECTION SYSTEM ARCHITECTURE

Detection of duplicate record is the process of identifying unique real world object from multiple or different records. So standardization of data is very important step in detection of duplications. Standardization converts the data in particular or specific standardize format [8,9,13]. For e.g. Pin code of any place is in 6 letters. So consider pin of Pune region is 400001 and any data shows the pin code of Pune is 00400001 then standardization perform on this data and convert it into the 400001. That means any string which have number of zero's before any non-zero numbers refer as a null.

Key Generation is every important and necessary task in detection of duplication. Key is selected as per categories of dataset. Duplication detection algorithms in this step various algorithms are compared which are based on blocking and windowing methods. Then compare the result of each algorithm.

Blocks of duplicate which are generating from algorithm evaluation are stored in result database, and statistical evaluation of this duplicates are perform. In statistical evaluation Completeness, Reduction Ratio and F-score are measure [4],[7],[12].

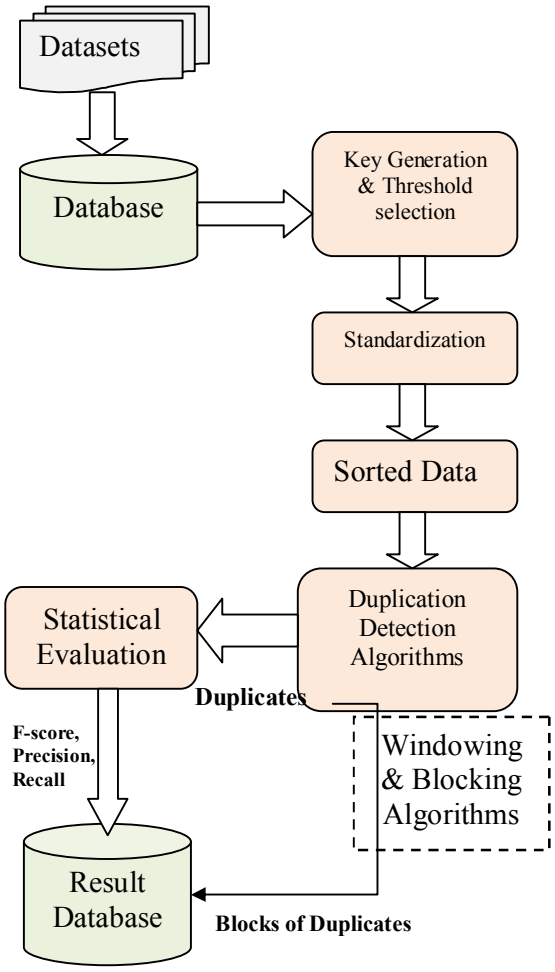


Figure 3. General Duplication Detection System Architecture

IV. CONCLUSIONS

The efficient Detection of duplicate records is challenging work because database cleaning is very complicated process. In this review paper we discuss various algorithms of windowing and blocking. The main focus of this paper is maintained efficiency and effectiveness of detection of duplicate records in large amount of databases. The adaptive size of window is useful than the fixed size window. Sorted block method provide good results in some cases than the adaptive windowing.

V. ACKNOWLEDGMENT

The authors would like to thank Shri Chairman Groups and Management and the Director/Principal Dr.Uttam Kalawane, Colleague of the Department of Computer Engineering and Colleagues of the varies Department the D.Y.Patil School of Engineering and Technology, Pune Dist. Pune Maharashtra, India, for their support, suggestions and encouragement.

REFERENCES

[1] M.Rehman, V.Esichaikul, "*Duplicate Record Detection For Database Cleansing*", Second International Conference on Machine Vision, 2009.

[2] S. Ong, A.Pei, "*A Comparative Study of Record Matching Algorithms*", Germany and University of Edinburgh, Scotland, 2008.

[3] M.Bilenko, B.Kamath, R.Mooney, "*Adaptive Blocking: Learning to Scale Up Record Linkage*", In Proceedings of the Sixth IEEE International Conference on Data Mining (ICDM-06), Hong Kong, December 2006, pp. 87-96.

[4] U. Draisbach, F. Naumann, S. Szott, and O. Wonneberg, "*Adaptive windows for duplicate detection*," Hasso-Plattner-Institut fur Software system technikaner Universitat Potsdam, Tech. Rep. 49, 2011.

[5] K.Prasad, S. Chaturvedi, T. Faruquie, L. Subramaniam, "*Automated Selection of Blocking Columns for Record Linkage*", IEEE,2012.

[6] J. Nin, V. Mulero, N.Bazan, Josep-L. L.Pey, "*On the Use of Semantic Blocking Techniques for Data Cleansing and Integration*",11th International Database Engineering and Applications Symposium, 2007.

[7] U. Draisbach and F. Naumann, "*A comparison and generalization of blocking and windowing algorithms for duplicate detection*," in Proceedings of the International Workshop on Quality in Databases (QDB),2009.

[8] Ahmed K. Elmagarmid, Panagiotis G. Ipeirotis, and Vassilios S. Verykios. "*Duplicate record detection: A survey*",IEEE Transactions on Knowledge and Data Engineering (TKDE), 19, 2007.

[9] D.Bharambe, S.Jain, A.Jain, "*A Survey: Detection of Duplicate Record*", International Journal of Emerging Technology and Advanced Engineering, Volume 2, Issue 11, November 2012.

[10] U. Draisbach, F. Naumann, S. Szott, O. Wonneberg, "*Adaptive Windows for Duplicate Detection* " , ACM SIGKDD interntional conference on Knowledge discover and data mining, NY, USA, 2011

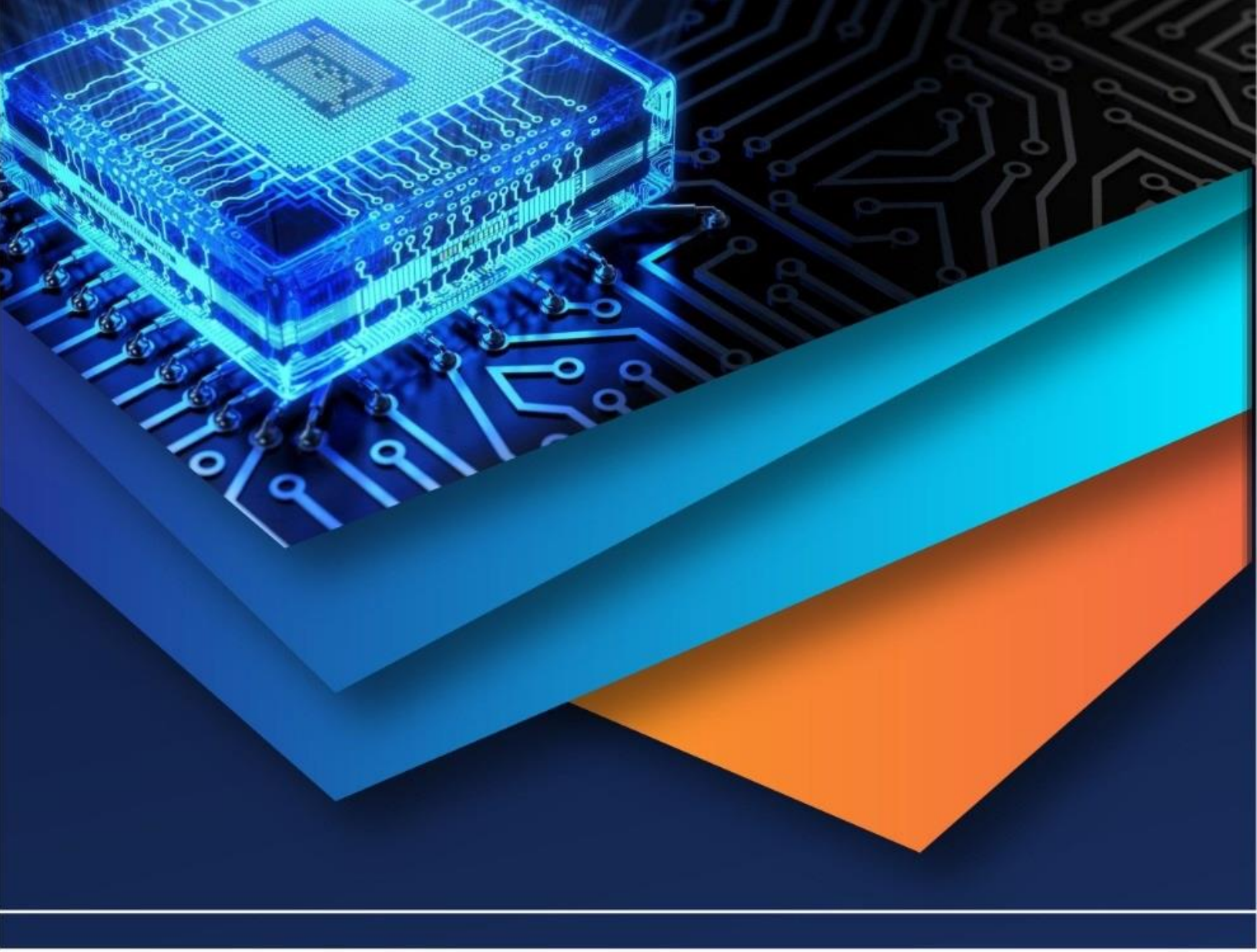
[11] S. Yan, D. Lee, M.-Y. Kan, and L. C. Giles, "*Adaptive sorted neighborhood methods for efficient record linkage,*" in Proceedings of the ACM/IEEE-CS joint conference on Digital libraries (JCDL), 2007, pp.185–194.

[12] R. Baxter and P Christen. , *"A comparison of fast blocking methods for record linkage,"* In In ACM SIGKDD workshop on Data Cleansing, Record Linkage and Object Consolidation, pages 25-27, Washington DC, 2003.

[13] Mauricio A.Hernandez, "*A Generalization of Band Joins and The Merge Purge Problem"*, Thesis Proposal, Department of Computer Science Columbia University New York, February 1995.

[14] M. Bilenko and R. J. Mooney, *"Adaptive duplicate detection using learnable string similarity measures,"* in Proceedings of the International Conference on Knowledge Discovery and Data Mining (KDD), 2003, pp.39–48.

[15] E. Rahm and H. Hai Do, *"Data Cleaning: Problems and Current Approaches"*, Bulletin of the IEEE Computer Society Technical Committee on Data Engineering.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY