ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 3, March 2019

A Review on Unsupervised Record Deduplication in Data Warehouse Environment

Keshav Unde, Dhiraj Wadhawa, Bhushan Badave, Aditya Shete, Vaishali Wangikar

School of Computer Engineering, MIT Academy of Engineering, Pune, Alandi, Maharashtra, India

ABSTRACT: Duplicate record detection process is known as the process of identifying pairs of records in dataset that corresponds to the same real world entity. Duplicate records significantly increase the cost for saving and managing data also its retrieval is delayed. To identify duplicate entries from dataset indexing techniques are used. These indexing techniques reduce complexities of deduplication process. For making indexing and ultimately deduplication process efficient, correct token (blocking key) formation is inevitable. Identification of appropriate token for record deduplication using an unsupervised method for learning blocking schemes is proposed in this paper.

KEYWORDS: Deduplication, Blocking, Indexing, Duplicate detection, Feature Selection, Disjunctive Blocking Scheme

I. INTRODUCTION

In a duplicate record detection process,the safe way to find all duplicates is to compare each element (record) with each other element. This approach is computationally expensive because the vast majority of comparisons are performed on two totally different records that have little to nothing in common. In general, pair selection divides the comparison space into either overlapping or non-overlapping partitions and performs the comparisons only within those partitions. One well-known, non-overlapping technique for pair selection is blocking. Blocking methods mitigate full pairwise comparisons by selecting a small subset of pairs from the database that are considered to be good candidates for pairwise comparison, while discarding the vast majority of pairs that are clearly non-coreferent. An unsupervised method for learning blocking schemes which is further used in record deduplication is discussed in this paper. The method consists of two phases. In the first phase, the algorithm efficiently generates a weakly labeled training set. In the second phase, the problem of learning blocking schemes from this weakly labeled set casted as a feature selection problem. Two main goals to achieve are,finding the best blocking key for a given dataset and creating an autonomous system (a service) for the entire duplicate detection process, with no need for human configuration. Objectives are -
1.To create a technique to automatically choose blocking keys that are both a good estimation of the record similarity and do not create too large partitions.
2.Enhancing efficiency of duplicate detection process.
3.To reduce the computation time required for duplicate records detection.

II. RELATED WORK

Variations of different indexing techniques such as Q-gram based indexing, suffix array based indexing, canopy clustering, sorted neighborhood indexing are surveyed by P. Christen

[1]. Their complexity is analyzed and their performance and scalability is evaluated within an experimental framework using both synthetic and real data sets.The number of candidate record pairs generated by these techniques has been estimated theoretically.

Elmagarmid et al. presented a comprehensive survey of the existing techniques used for duplicate records detection

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 3, March 2019

[2]. They covered similarity metrics that are commonly used to detect similar field entries and presented duplicate detection algorithms that can detect approximately duplicate records in a database. They also covered multiple techniques for improving the efficiency and scalability of approximate duplicate detection algorithms.

Two competing approaches for selecting promising record pairs for comparison are blocking and windowing. Blocking methods partition records into disjoint subsets, while windowing methods, in particular the Sorted Neighborhood Method, slide a window over the sorted records and compare records only within the window. Draisbach et al. proposed new algorithm called Sorted Blocks in several variants, which generalizes both approaches [3].

Partitioning technique is generally used to avoid many unnecessary comparisons. However, partitioning keys are often poorly chosen. Vogel et al. proposed a technique to find suitable blocking keys automatically. Blocking keys are created based on unigrams [4]. Blocking key creation is accompanied with several comprehensive experiments on large artificial and real-world datasets. Using this approach, high quality blocking keys on a given training dataset can be discovered.

Naumann et al. presented a novel approach called one-to-some or 1:k assignment to assign similarity measures to attributes for supporting duplicate detection process [5]. The approach features are minimal required user interaction and self-configuration for the provided input data. With this matching approach,the most relevant attributes can be identified and appropriate similarity measures can be derived.

Bilenko et al. introduced an adaptive framework for automatically learning blocking functions that are efficient and accurate.They described two predicate-based formulations of learnable blocking functions and provide learning algorithms for training them [6]. There are two types of blocking functions, disjunctions of blocking predicates and predicates combined in disjunctive normal form (DNF).

Michelson et al. proposed a machine learning approach to automatically learn effective blocking schemes [7]. The proposed blocking scheme fulfills the two main goals of blocking, maximizing both pairs completeness and reduction ratio. It can reduce the candidate matches for record linkage without losing matches.

Indexing is an important step in the deduplication process for reducing the search space by bringing similar records closer to each other using a blocking key criterion. To make the deduplication process less dependent on human domain knowledge, automatic selection of optimal blocking keys is necessary. Ramadan et al. proposed an unsupervised learning technique that automatically selects optimal blocking keys for building indexes that can be used in deduplication

[8]. They used multiple keys with multi-pass sorted neighborhood, one of the most efficient and widely used indexing techniques for deduplication. This technique gives optimal blocking/sorting keys which significantly increase the efficiency of deduplication process while maintaining the quality of matching results.

Wangikar et al. studied and implemented Sorted Neighborhood based de-duplication techniques [9].Also during implementation Adaptive and Non-Adaptive Sorted Neighborhood Methods are experimented and validated.Accumulative Adaptive SNM (AASNM), Incrementally Adaptive SNM (IASNM) are adaptive versions of SNM while Duplicate Count Strategy (DCS) is a Non Adaptive SNM. A Group based Accumulative Adaptive Method (GAASNM) is proposed to minimize the record comparisons.

The main aim of indexing techniques is to minimize the number of data pairs by eliminating apparent non-matching data pairs by maintaining maximum quality of matching. Reddy et al. surveyed various indexing techniques to analyze complexity and evaluate scalability using various datasets

[10]. Measures such as reduction ratio, pairs completeness and pairs quality are used for evaluating performance.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 3, March 2019

III. PROBLEM STATEMENT

The problem of identifying approximately duplicate record in database is an essential step for data cleaning and data integration process. To implement the automated deduplication framework for identifying the duplicate records with the help of unsupervised blocking scheme using machine learning algorithm.

IV. EXISTING SYSTEM

Deduplication is well known challenge in data integration.Manually identifying duplicate record is troublesome for even moderate size database and almost impossible for large database.Also traditional techniques of identifying duplicates include comparing each entity with other entity to determine whether the two records refers to same entity which grows quadratically with input and is impractical for large database.Ad-hoc and domain dependent solution with human intervention still used for identifying duplicate records.

V. SYSTEM OVERVIEW

An indexing function takes field value from tuple and returns a set containing Blocking Key Values. An example of an indexing function is Tokens. Tokenizing process takes a field value of a tuple and parses it into a set of tokens using a set of common delimiters (such as whitespace and comma). This set is then returned as the Blocking Key Values set that identifies the blocks in which the tuple is placed.

General blocking predicates whose indexing functions are not associated with specific fields are then used. For example, general blocking predicate Contains Common Token can be applied to many fields in dataset. Whereas, a specific blocking predicates explicitly pairs a general blocking predicate to a specific field. A Disjunctive Blocking Scheme is merely a disjunction of specific blocking predicates while a general DNF Blocking Scheme is a disjunction of terms, where each term is a conjunction of specific blocking predicates. A blocking scheme may be applied to a tuple pair, which is said to be covered if the scheme returns True for the pair, just like with specific blocking predicates. The learning blocking schemes are useful for generating a weakly labeled training set. The user may specify a maximum of duplicates and non-duplicates to be returned. The Fisher discrimination criterion is then used to determine the best features using FisherScore Algorithm. The Fisher scores are used in the algorithm to impose an ordering in which eligible feature elements are considered. Finally Support Vector Machine classification algorithm applied for detecting duplicate records with the help of generated weakly labeled dataset.
Advantages are: 1. Using unsupervised blocking scheme in duplicate detection process improves overall accuracy rate. 2.Blocking scheme minimizes the number of candidate pairs. 3.Reducing the time consumption for duplicate record detection.

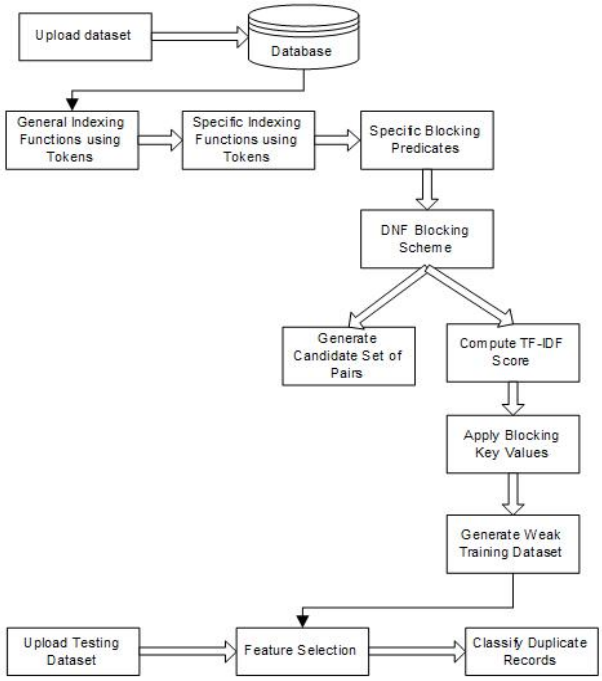
ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 3, March 2019



deduplication process without human intervention. Correct blocking key identification improves an efficiency of indexing and deduplication process.As a part of future work, we will try to identify more efficient algorithms and techniques for improving efficiency of overall deduplication process.

VI. CONCLUSION

Blocking technique used for deduplication significantly reduced comparisons as compared to full pairwise comparison. An unsupervised way for learning good blocking schemes is proposed in this paper. The proposed algorithm found to present favorable results compared to a supervised state-of-the-art algorithm. Automatic token formation facilitates

REFERENCES

[1] P. Christen, A survey of indexing techniques for scalable record linkage and deduplication, Knowledge and Data Engineering, IEEE Transactions on, vol. 24, no. 9, pp. 15371555, 2012.
[2] A. K. Elmagarmid et al. ,Duplicate record detection: A survey, Knowledge and Data Engineering, IEEE Transactions on, vol. 19, no. 1, pp. 116, 2007.
[3] U. Draisbach et al. ,A generalization of blocking and windowing algorithms for duplicate detection. In Proceedings of the International Conference on Data and Knowledge Engineering (ICDKE), 2011.
[4] T. Vogel et al. ,Automatic Blocking Key Selection for Duplicate Detection based on Unigram Combinations, In proceedings of the International workshop on Quality in databases (QDB), 2012.
[5] F. Naumann et al. ,Instance-based one-to-some assignment of similarity measures to attributes. In Proceedings of the International Conference on Cooperative Information Systems (CoopIS), 2011.
[6] M. Bilenko et al., Adaptive blocking: Learning to scale up record linkage, in Data Mining, 2006. ICDM06. Sixth International Conference on. IEEE, 2006, pp. 8796.
[7] M. Michelson et al. ,Learning blocking schemes for record linkage, in Proceedings of the National Conference on Artificial Intelligence, 2006, p. 440.
[8] Ramadan et al. ,Unsupervised Blocking Key Selection for Real-Time Entity Resolution, In Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'15), 2015.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 3, March 2019

[9] V. Wangikar et al. ,Study and Implementation of Record De-duplication Algorithms ,In Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies ,2016.

[10] S.Reddy et al. , Complexity Analysis of Indexing Techniques for Scalable Deduplication and Data Linkage, In proceedings of International Journal of Engineering Research and Applications, 2014.

[11]M. Kejriwal et al. ,An Unsupervised Algorithm for Learning Blocking Schemes, In Proceedings of 13th International Conference on Data Mining (ICDM), 2013 .