ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 1, Issue 8, October 2012

A Detailed Survey on Various Record Deduplication Methods

Lalitha. L[1], Maheswari.B[2],Dr.Karthik.S[3]

***Abstract*— Deduplication is the key operation in data integration from multiple data sources. To achieve higher quality information and more simplified data representation, data preprocessing is required. Data cleaning is one among the data preprocessing steps. Data cleaning includes the process of parsing, data transformation, duplicate elimination and statistical methods. If two records represent the same real world entity then it is called duplicated records. The problem of detecting and eliminating duplicate records is called record deduplication. This paper presents an analysis of record deduplication techniques and algorithms that detect and remove the duplicate records.**

***Index Terms*—Deduplication, Data cleaning, Data preprocessing, Record Linkage, Record matching.**

I. INTRODUCTION

Data mining is the extraction of hidden predictive information from large databases. It is a new powerful technology with great potential to help companies focus on the most important information in their data warehouses. In real world, data mining technique is applicable to many areas like banking systems, educational systems, airline reservation systems etc. Digital media has become a challenging problem for data administrators when volume of information is increased. Built on data gathered from different sources, data repositories such as those used by digital libraries and e-commerce brokers may present records with disparate structure [1]. Today's IT based economy, databases play an important role.

When integrating data from different sources for implementing a data warehouses, organizations become aware of potential systematic differences or conflicts [2]. At the time of integrating data from multiple heterogeneous sources, record replicas and duplicates will occur. In a data repository, a record that refers to the same real world entity or object is referred as duplicate records. Due to the duplicate records and dirty data, many problems will occur like performance degradation, quality loss and increasing operational costs [1]. For avoiding these types of problems, data preprocessing steps are performed. The preprocessing steps are data cleaning, data integration, data transformation and data reduction [3].

Data cleaning is the process of detecting and correcting in accurate records from a record set, table or data base. After cleaning, a data set will be consistent with other similar data sets in the systems. The inconsistencies detected or removed are originally caused by user entry errors or corruption in transmission. Data cleaning differs from data validation means data is rejected from the system at entry and is performed at entry time rather than batches of data [4]. Data cleaning includes the process of parsing, data transformation, duplicate elimination and statistical methods [4]. Record deduplication is the process of identifying and removing duplicate entries in a repository. It is also referred as data cleaning, record linkage and matching [1].

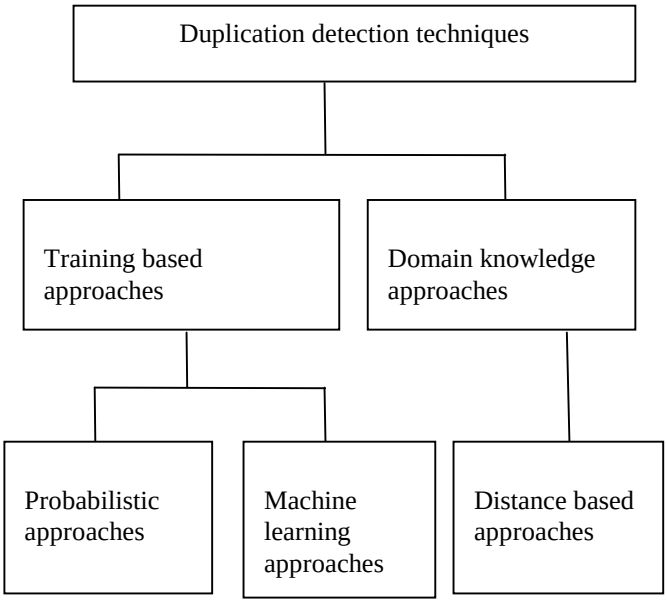

Fig. 1: Methods for duplicate record detection

II.RELATED WORK

A. PSO Algorithm Based Deduplication

K. Deepa et al. [5] proposed a heuristic global optimization method called Particle Swarm Optimization algorithm for record deduplication. They considered the fitness function of the PSO algorithm and it is based on swarm of data. Here the proposed approach has two phases such as training phase and duplicate detection phase. First they find the similarity between the all attributes of record pairs using Levenshtein distance and cosine similarity. Then they formed the feature vectors for representing the set of elements which required detection of duplicates. From this feature vectors, they found the duplicate records by using the PSO algorithm. PSO algorithm is very simple and it needs

Manuscript received Sep 15, 2012.
Lalitha.L[1], II/ME (CSE), SNS College of Technology, Coimbatore, (e-mail: lalithamyt@gmail.com).
Maheswari.B[2], AP/CSE, SNS College of Technology, Coimbatore, (e-mail: maheswari.bk@gmail.com).
Dr.Karthik.S[3], Dean/CSE, SNS College of Technology, Coimbatore. (e-mail: profskarthik@gmail.com).

fewer parameters. This algorithm has no overlapping and mutation calculation.
It provides more accuracy in record deduplication than genetic algorithm. Even though PSO has some advantages, it cannot work out the problems of scattering and optimization and cannot work out the moving rules of particles in the energy field [6].

B. Active Learning Based Deduplication

Sunita Sarawagi and Anuradha Bhamidipaty [7] proposed an interactive learning based deduplication system called Active Learning led Interactive Alias Suppression (ALIAS). This technique automatically constructs the deduplication function by interactively finding the challenging training pairs. An active learner actively picks the subset of instances. It easing the deduplication task by limiting the manual effort for inputting simple, domain specific attributes similarity functions. It interactively labeling a small number of record pairs. First they took the small subset of pair of records. Then they find the similarity between records and this initial set of labeled data creates the training data for the preliminary classifier. To improve the accuracy of classifier they selected only n instances from the pool of unlabeled data [7]. They conclude that, active learning process is practical effective and provide interactive response to the user. It is easy to interpret and efficient to apply on large datasets. Active learning requires some training data but in some real world problems the training data are not available, so active learning technique is not suitable for all the problems [2].

C. Divide and Conquer Based Deduplication

Bilal Khan et al. [8] suggested an approach for duplicate record detection and removal. In this approach, they first convert the attributes of data into numeric form. Then, this numeric form is used to create clusters by using K-Means clustering algorithm. The use of clustering reduces the number of comparisons. After that the divide and conquer technique is used in parallel with these clusters for identification and removal of duplicated records. Here, this technique identifies all type of duplicated records like fully duplicated records, erroneous duplicated records and partially duplicated records. This technique is only applicable for single table instead of multiple sorted tables. The performance is measured by using the terms like true positives, false positives, false negatives, precision, recall and F-Score [8].

D. Indexing Based Record Deduplication and Record Linkage

Peter Christen [9] surveyed various indexing techniques for record linkage and dedupliaction. Record linkage refers to the task of identifying records in a data set that refers to the same entity across different data sources [10]. Blocking technique is used in traditional record linkage approach. Blocking key values are used to place the records into different blocks. According to this BKV, the matched records are placed in same block and non matching records into different blocks. The record linkage process has divided into two phases: Build and Retrieve. In build phase, at the time of linking two data bases, a new data structure is formed: i) Separate index data structures ii) Single data structures with common key values. The hash table data structure is also used for indexing. In retrieve phase, the retrieval of records from block and it will be paired with other records which having same index value. This resulting vector given to classification steps. There are many indexing techniques available. There are as follows.

1) Traditional Blocking: In this technique, placing the records in wrong block may occur and total number of record pairs to be generated is not predictable [9].

2) Sorted Neighborhood Indexing: Sorting the database according to BKV is the main idea behind in this indexing. A window is used for generation of record pairs. One drawback is occurred when the small window size is chosen because it is not enough for large number of records [9].

3) Q-Gram Based Indexing: The main goal of this technique is to index the databases that have the similar records, BKV inserted into same block. This method was implemented within a relational database and using SQL statements [9].

4) Suffix Array Based Indexing: The main idea in this indexing is insert the BKV_s and their suffixes into a suffix array-based inverted index. It contains strings or sequences and their suffixes in an alphabetical sorted order. This indexing is mainly suitable for the larger blocks [9].

5) Canopy Clustering: Here clustering technique is used. Overlapping clusters are called canopies. These clusters can be created using thresholds or nearest neighbors. Canopy clustering increases the number of true matches than any other indexing techniques.

6) String-Map Based Indexing: The idea is to first map records into a multi-dimensional space and by a mapping into a second lower-dimensional metric space where edit-distance calculations are performed. All the strings are mapped into a multidimensional space using R-tree. Clusters of similar objects are retrieved using indexing approach. In this technique efficient matching is obtained.

These techniques are mainly used to reduce the number comparison between the records. This can be achieved by removing non matching pairs from the block [9].

E. Unsupervised Duplicate Detection (UDD) Algorithm

Weifeng Su et al. [11] proposed an unsupervised, online record matching method called Unsupervised Duplicate Detection (UDD) algorithm. There are two classifiers in UDD for iteratively identify the duplicate records. The duplicate records from the same source are removed using the exact matching method. In this method relative distance of each field of the records are calculated and according to this value, field's weight will be assigned. After that Weighted Component Similarity Summing (WCSS) Classifier utilizes this weight set for matching the records from various data sources. It places the duplicate records in the positive set and non duplicate records in the negative set. The SVM classifier again identifies duplicates from the positive set. These two classifiers iteratively working together and identify the duplicate records in

efficient manner. The iteration stops when new duplicates cannot be found. This algorithm mainly used in the web databases because UDD does not require human labeled-training data from the user. So it solves the online duplicate detection problem where the query results are generated on-the-fly [11].

In SVM based record deduplication only the concrete implementation has been done. However sometimes it requires an initial approximated training set to assign weight [12].

F. Removing Fuzzy Duplicate Records Using an Adaptive Framework

Hamid Haidarian Shahri and Saied Haidarian Shahri [13] invented an adaptive and extensible framework for eliminating the duplicates. In this framework there are six steps of workflow for duplicate elimination. In all these steps, user can select appropriate items based on that step. In first step, a clustering algorithm is selected for grouping the records(duplicates).In second step, attributes of records are selected for comparing a pair of tuples. In the next step, similarity functions are selected for measuring attribute similarity. In the fourth step, fuzzy rules are used in the fuzzy inference engine to detect the duplicates. For that, it uses the rule viewer, logging membership functions and machine learning capabilities. Neuro-fuzzy modeling is used for applying the learning technique. By using the rule viewer the user can fine-tune the system's rules and membership functions. For this tuning Adaptive Network-based Fuzzy Inference system (ANFIS) is used in this framework. In the fifth step, membership functions selection is done. At last step, the selection of merging technique is done for choosing which tuple will be the prime representative of the duplicates. In this way the duplicate elimination process is done in this framework. Their framework provides the fuzzy logic and removes the hard coding for duplicate elimination. Another advantage is the inclusion of machine learning capabilities. They found that the efficiency of this framework will be improved by adding smarter, more sophisticated domain-dependent similarity functions [13].

G. Deduplication Using Febrl System

Peter Christen [14] provided the overview about the Febrl system. Febrl system (Freely Extensible Biomedical Record Linkage) is an open-source data cleaning toolkit. It has two components: first one manages the data standardization using Hidden-Markov Models (HMM_s) and second one performs the actual duplicate detection. Febrl requires the training to correctly parse the database entries. It implements the variety of string similarity metrics. Febrl uses the phonetic encoding to detect similar names [2]. Febrl provides the graphical user interface that helps the end user who do not have programming experience. Even though Febrl has GUI, it also has some limitations like poor scalability, slowness, unclear error messages and complex installation. It requires large amount of memory for large data sets. In future work, additional output options should be added to Febrl that allow flexible merging of the linked records into a linked output data set. Additional field comparison, classification and indexing methods are needed to improve the performance of the Febrl. Febrl's performance should be improved for reducing the amount of memory required when deduplicating or linking larger data sets [14].

III.CONCLUSION

An analysis of the existing record deduplication techniques and frameworks is done here. Deduplication and record linkage is a crucial step in data integration. From this survey, it is possible to conclude that the existing algorithms require more memory for deduplication. It is also time consuming process. In future a deduplication algorithm can be designed for reducing the number of comparison between the records such that it reduces time consumption and utilizes less memory space.

REFERENCES

[1] Moises G. de Carvalho, Alberto H.F. Laender, Marcos AndreGoncalves, and Altigran S. da silva, "A Genetic Programming Approach to Record Deduplication", *IEEE Trans. Knowledge and Data Eng.*, vol. 24,no. 3, pp. 399-412, Mar. 2012.

[2] A.K. Elmagarmid, P.G. Ipeirotis, and V.S. Verykios, "Duplicate Record Detection: A Survey", *IEEE Trans. Knowledge and Data Eng.*, vol. 19, no. 1, pp. 1-16, Jan. 2007.

[3] V. Subramaniyaswamy, S. Chenthur Pandian, "A Complete Survey of Duplicate Record Detection Using Data Mining Techniques", *Information Technology Journal 11(8)., ISSN 1812-5638,* pp.941-945, 2012.

[4] "Data Cleaning", http://en.wikipedia.org/wiki/Data_cleaning

[5] K. Deepa, R. Rangarajan, "Record Deduplication using Particle Swarm Optimization", *European Journal of Scientific Research ISSN 1450-216X.*,vol.80,no. 3, pp. 366-378, 2012.

[6] Qinghai Bai, "Analysis of Particle Swarm Optimization Algorithm", *Computer and Information Science*, vol.3, no.1, pp. 180-184, Feb. 2010. www.ccsenet.org/cis.

[7] S. Sarawagi and A. Bhamidipaty, "Interactive Deduplication Using Active Learning", *Proc. Eighth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining(KDD'02),* pp.269-278, 2002.

[8] Bilal Khan, Azhar Rauf, Sajid H. Shah and Shah Khusro, "Identification and Removal of Duplicated Records", *World Applied Sciences Journal 13(5): ISSN 1818-4952*, pp.1178-1184, 2011.

[9]Peter Christen, "A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication", *IEEE Trans. Knowledge and Data Eng.*, vol. 24, no. 9, pp. 1537-1555, Sept.2012.

[10] "Record Linkage", http://en.wikipedia.org/wiki/Record _linkage

[11]Weifeng su, Jiying Wang, Frederick H. Lochovsky, " Record Matching over Query Results from Multiple Web Databases", *IEEE Trans. Knowledge and Data Eng.*, vol. 22, no. 4, pp.578-588, April. 2010.

[12] A.Faritha Banu, C.Chandrasekar,"A Survey on Deduplication Methods", *International Journal of Computer Trends and Technology, ISSN: 2231-2803*, vol. 3, Issue. 3, pp.364-368, 2012, http://www.internationaljournalssrg.org

[13] Hamid Haidarian Shahri, Saied Haidarian Shahri, "Eliminating Duplicates in information Integration: An Adaptive, Extensible Framework", *IEEE Computer Society 1541-1672,* pp. 63-71, September/October 2006.

[14] Peter Christen, Development and User Experiences of an Open Source Data Cleaning, Deduplication and Record Linkage System", *SIGKDD Explorations.*, vol. 11, Issue 1, pp. 39-48.

[15] V.P.Arunachalam,S.Karthik, "A Novel approach for mining inter-transaction itemsets", European Scientific Journal, 8(14).

Lalitha.L received Diploma in Computer Technology at M.P.Nachimuthu M.Jaganathan Polytechnic College, Chennimalai in 2002 and B.E degree in Computer Science and Engineering at M.P.Nachimuthu M.Jaganathan Engineering College , Chennimalai in 2011. She is currently pursuing M.E degree in Computer Science and Engineering at SNS College of Technology, Coimbatore. Her areas of interests are Networks, data warehousing and data mining.

Ms.B.Maheswari is presently Assistant professor, Department of Computer science and Engineering, SNS college of Technology, affiliated to Anna University- Coimbatore, Tamil Nadu, India. She received her Bachelor of Engineering degree in Computer Science and Engineering in 2007 and received her Master of Computer Science and Engineering degree in SNS College of Technology, Coimbatore, India. Her research interests include network security, Mobile Computing and wireless systems. She has published more number of articles in International/ National Journals and Conferences.

Professor Dr.S.Karthik is presently Professor & Dean in the Department of Computer Science & Engineering, SNS College of Technology, affiliated to Anna University- Coimbatore, Tamilnadu, India. He received the M.E degree from the Anna University Chennai and Ph.D degree from Ann University of Technology, Coimbatore. His research interests include network security, web services and wireless systems. In particular, he is currently working in a research group developing new Internet security architectures and active defense systems against DDoS attacks. Dr.S.Karthik published more than 35 papers in refereed international journals and 25 papers in conferences and has been involved many international conferences as Technical Chair and tutorial presenter. He is an active member of IEEE, ISTE, IAENG, IACSIT and Indian Computer Society.

.