ISSN (Online): 2230-7893
www.IJCEM.org

A Signature-based Clustering Approach for Record Matching from Multiple Databases

Prabhjyot Juneja[1] and Urjita Thakar[2]

[1,2] Department of Computer Engineering, Shri G.S. Institute Of Technology and Science, Indore, Madhya Pradesh, India

Abstract

Various business analytics depend on collection and analysis of large volumes of data. Data from multiple sources may need to be integrated and examined to draw fruitful results. Lack of a global unique identifier and other data inconsistencies pose challenges in linking such data. Existing record linkage approaches set certain prerequisites in form of either training data or known match status of databases under consideration. These approaches rely particularly on the problem datasets and are practically infeasible for large volumes of disparate real-world data. The proposed signature-based clustering approach generates multiple signatures of each record and clusters similar records together using affinity propagation clustering technique. Each set of clustered records is then represented by its representative feature.

High performance of the proposed algorithm on various real-world and synthetic datasets indicates its efficiency over existing approaches. The proposed system generates matching results with an average 92% accuracy. Representation of each set of similar records by a single representative feature greatly reduces the number of overall comparisons and memory requirements of the matching process, thereby improving the matching quality.
Keywords: record matching, multiple databases, clustering, signatures, business analytics.

1. Introduction

Record matching is the procedure of identifying records in same or different databases that refer to the same real-world entity. Finding records that refer to the same entity within a single database is commonly known as duplicate detection [1]. From decades, record matching has been used to link records within two similar databases. It finds applications in matching census data [2], product descriptions, bibliographic databases, etc. With the boost of innovation and advancement in technology, this technique has been of interest in mining data for e-commerce applications like customer profiling, predictive analysis and in other domains like entity disambiguation [3], natural language processing, etc.

A large number of applications and web services collect users' personal and social data and utilize them for drawing significant inferences. Absence of a unique identifier is a major problem faced in these tasks. As a result, matching needs to be performed using the available personal information. Moreover, multiple databases comprise of different features, inconsistent data formats, missing and erroneous data. In addition to this, in many applications, there is no ground truth data or 'known match status' to measure the quality of matching results. Missing potential matches and deriving false matches, both incur high costs. All these factors make record matching from multiple databases a challenging task.

Previous works proposed solutions for deduplication in a single database or record matching in two databases. No significant work has been done in the area of record matching from multiple databases. Data driven approaches [4] for performing record linkage from disparate data sources are based on matching rules derived from supervised learning techniques [5,6]. However, these approaches require training data, thereby making the solution specific to the problem dataset.

In this paper, an algorithm is proposed to identify similar records through clustering. Data features are utilized to create signatures for each record and perform record-matching. Since each record in a database will ideally be represented by one cluster, the number of clusters is unknown at the beginning of the process. Also the number of clusters can be very large when the databases that are matched contain many records. However, the clusters generated in record matching can be very small, containing only a few records. A representative feature is selected to represent each set of clustered records. If a new record does not match any representative feature, a new cluster is created for that record. This approach results in improving the matching quality to a great extent.

Rest of the paper is organized as follows. In the next section, related work done in this field is discussed. In section 3, proposed methodology is discussed in detail.

Observations and results derived from experimentation are detailed in sections 4 and 5. The paper is concluded in Section 6.

2. Related Work

In this section, work done in the domain of record matching has been discussed. Previous researches are based on supervised learning, semi-supervised learning, active learning and unsupervised learning approaches [1]. Only unsupervised learning approaches do not pose a requirement of training data to perform matching. As stated earlier, data from such diverse range of applications generally lacks known match status, thus it is practically infeasible to generate a training dataset that covers all possible matching samples from each set of input databases.

Different unsupervised approaches based on clustering have been proposed by researchers. Some clustering techniques partition data objects into a fixed number of clusters while some approaches generate graphs corresponding to clusters or generate clusters that correspond to dense areas where many data objects are located close to each other.

Partitioning based clustering algorithms [7] partition records into a hierarchy of clusters. But these require the number of clusters to be specified at the beginning and thus are not applicable for clustering records in data matching applications. Monge [8] proposed an early clustering approach where records are clustered according to some similarity measure and a priority queue is kept in memory consisting of the most recently formed clusters. Initially all records to be matched were sorted on the basis of a sorting key. This approach reduced the number of record pair comparisons that incur during matching, but this approach has memory constraints and the results highly depend on the sorting key selected at the beginning.

Approaches that generate graphs corresponding to clusters [9] employ pair-wise comparison and classification technique. A drawback of these approaches is that the minimum similarity threshold that has been used to categorise pairs of records into matches and non-matches determines the structure of the cluster graph. This threshold is a global parameter applied to all compared record pairs.
Single key blocking (SKB) based and Composite key blocking (CKB) based record linkage approaches [10] group records on the basis of a single field and two fields respectively. These approaches decrease the number of record comparisons but reduce the matching accuracy by missing potential matches.

A generalized algorithm is therefore needed to perform record matching by effectively using the available data features.

3. Proposed Methodology

In the proposed system, preprocessing is performed to ensure that datasets are in the same format and to reduce their dimensionality. The prepared datasets are merged to form a single set. The merged dataset is then subjected to signature-based clustering. For each record, a signature is generated comprising of its relevant attributes. To evaluate this method, different signatures are generated comprising of: (person's first name and last name), (person's last name, first name and date of birth), (person's last name, father's name, first name and date of birth). These signatures are then subjected to Affinity propagation clustering [11]. Similar signatures, corresponding to similar records are clustered together as illustrated in figure 1.

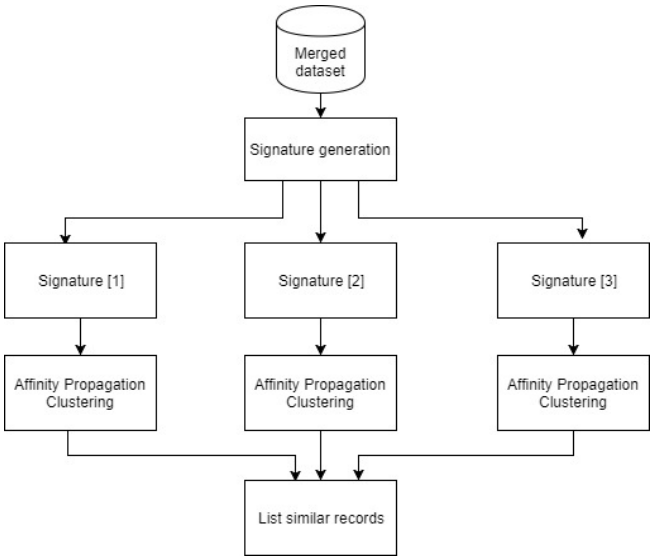

Fig 1. Signature based clustering

The components of the system are discussed further.

3.1 Data Preprocessing

It is done to ensure that datasets are in same format and to reduce their dimensionality. It consists of the following tasks:

- Data cleaning and munging
- Standardization and tokenisation
- Segmentation into output fields

Data cleaning aims at removal of unwanted characters and tokens. Data munging involves filling in missing values and extracting salutations from names. The objective of segmentation is to have each output field contain a single piece of information, made of one or a small number of tokens, rather than having several pieces of information in one field or attribute. The values in these output fields are then used in the detailed comparison of record pairs, which leads to much improved matching quality. Rule-based segmentation technique is used for segmentation of different types of input data, such as personal names, business names, or addresses. A new record identifier is created for each record to identify its source database. The prepared datasets are then subjected to integration.

3.2 Integration of Datasets

Records from different datasets are integrated together so that further computations are performed on a single set. Datasets consisting of varying schema can be merged if there is atleast one common attribute. Datasets with exactly similar schema can be integrated altogether. Since the prepared datasets comprise of attributes pertaining to personal information, they are merged on these common features.

3.3 Signature Generation

In this step, the attributes corresponding to personal information, such as a user's First name, Last name, Full name and date of birth are used to create signatures. Different signatures are created for each record to trace optimum matching results.

3.4 Affinity Propagation Clustering

The signatures are treated as data samples and are clustered on the basis of jaccard distance metric [9]. Similar signatures, corresponding to similar records are clustered together by affinity propagation clustering [11].

In this method, clusters are created by exchanging messages between pairs of data points until convergence. On the basis of the messages exchanged, an exemplar or representative feature is selected for each cluster. This approach does not enforce equal-sized clusters, and the number of clusters are selected on the basis of the data provided. For this purpose, the two input parameters are: preference and damping factor [11].

Records are clustered such that records within a cluster are more closely related to one another than the records assigned to different clusters. It has been observed that, few records might contain flipped values of first name and last name, or flipped values of address components. The traditional method of field comparison compares values only in respective fields of records. But, in the designed system, to capture such records, different combinations of signatures are clustered.

The results obtained by experimentation on different datasets are discussed in the following section.

4. Experimentation Results

4.1 Data Description

To demonstrate the efficiency of proposed approach, it was tested on real world datasets and synthetic datasets comprising of personal information of people. Real world datasets include Jansunvai and CM teerth yatra databases and synthetic datasets were generated from data generators. The details of all datasets are as under.

Table 1: Description of Datasets

Dataset	Instances	Features	Matching Records
Jansunvai	5000	11	Unknown
CM teerth yatra	5922	19	Unknown
Synthetic dataset1	1000	11	500
Synthetic dataset2	2500	11	1500
Synthetic dataset3	5000	11	1000

4.2 Evaluation of Obtained Results

Real datasets are merged to form a single set, whereas synthetic datasets are merged to form another set. Results obtained by the proposed approach vary on the basis of signature selected for clustering and the value of damping factor. The number of clusters obtained in each case are shown in fig. 2 and 3.

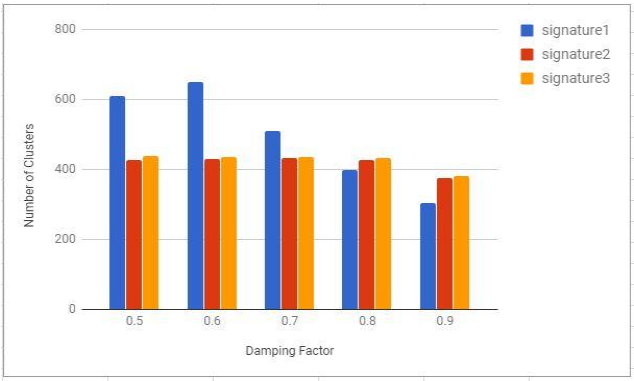

Fig.2 Performance of proposed approach on integrated real dataset

Here, signatures represented correspond to:

- signature1 : Applicant's first name, last name
- signature2 : Applicant's last name, first name and date of birth
- signature3 : Applicant's last name, father's name, first name and date of birth

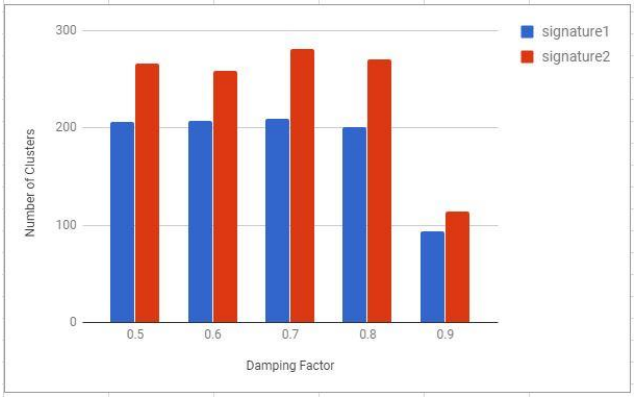

Fig.3 Performance of proposed approach on integrated synthetic dataset

Here, signatures represented correspond to:

- signature1 : Person's first name, last name
- signature2: Person's last name, first name and date of birth.

From the results depicted in fig.2 and fig3, it can be deduced that maximum value of damping factor results in lower number of clusters, which in turn results in merging of distinct clusters together. Close to accurate results are obtained on damping factor 0.6 and signature1 in real datasets and signature2 in synthetic datasets. Thus, selection of signature is a crucial factor that affects the performance of the system.

The proposed approach is compared to the existing Single key blocking based indexing approach (SKB) and Composite key blocking based indexing approach (CKB) for record linkage. In these approaches, name fields are encoded using Double metaphone phonetic encoding function [1]. Performance of these algorithms is evaluated on different key fields used for indexing and different signatures selected for clustering. For synthetic datasets, the number of matching records and the matching record pairs are known. Thus, the proposed algorithm is evaluated on these datasets. The results obtained are illustrated in Table 2.

Table 2: Evaluation metrics on synthetic datasets

Approach	Key Attributes	Precision	Recall	Accuracy	F-score
SKB based	Encoded Full name	0.94	0.7	0.84	0.81
CKB based	Encoded Firstname, DOB	0.97	0.67	0.83	0.79
Proposed approach	Signature1	0.83	**0.95**	**0.89**	**0.88**
Proposed approach	Signature2	0.93	**0.95**	**0.94**	**0.94**

It can be deduced from the results depicted in table 1 that the proposed signature-based clustering algorithm consistently finds greater number of matches than existing SKB and CKB record linkage approaches and thus, has a higher recall and f-score.

5. Discussion

Performance comparison with other approaches denotes that the proposed approach efficiently finds maximum similar records from multiple databases. Existing record linkage approaches generate pairs of similar records whereas the proposed approach generates clusters of similar records. This significantly reduces memory overhead incurred in saving all similar records in the database. Only the exemplars or representative features created for each cluster can be stored in memory. New records are compared to the representative features of each cluster, thereby decreasing the number of record comparisons to a great extent. Also, the designed system finds true matches with an average 92% accuracy. However, the results obtained vary according to the key fields selected for each dataset. Signature comprising of first name and last name attributes result in capturing more number of matches in the test datasets.

6. Conclusion

As data mining from multiple heterogenous sources is rapidly becoming common, effective, scalable and time-efficient matching techniques are necessary. In this work, an algorithm is proposed to identify similar records across multiple heterogeneous databases. Data features are utilized to create signatures and perform record-matching.

Thorough experimental evaluation on real world and synthetic datasets denotes that the proposed algorithm succeeds in tracing a large number of matching records with an average 92% accuracy and 95% recall, marking an

enhancement of 10% and 30% respectively over existing record linkage approaches.

The designed system performs equally well in the cases of insertion of new records and deletion or modification of the existing records in any of the input databases. The approach presented can be enhanced by use of leveraged affinity propagation clustering to make it scalable for large datasets.

References

[1] Ahmed K. Elmagarmid,Panagiotis G. Ipeirotis,Vassilios S. Verykios, "Duplicate Record Detection: A Survey", IEEE Transactions on Knowledge and Data Engineering, vol. 19, Issue:1, pp. 1-16, 2007.

[2] Gill, L., "Methods for automatic record matching and linking and their use in national statistics", Tech. Rep. Methodology Series, no. 25, National Statistics, London, 2001

[3] M. Bilenko, S. Basu and M. Saami, "Adaptive Product Normalization: Using Online Learning for Record Linkage in Comparison Shopping," in Proceedings of the Fifth IEEE International Conference on Data Mining, 2005.

[4] Colin Conrad, Naureen Ali, Vlado Keselj, Qigang Gao, "ELM: An Extended Logic Matching Method on Record Linkage Analysis of Disparate Databases for Profiling Data Mining", IEEE Conference on Business Informatics, 2016

[5] T. Joachims, "Making Large-Scale SVM Learning Practical", Advances in Kernel Methods—Support Vector Learning, MIT Press, 1999.

[6] A.E. Monge and C.P. Elkan, "An Efficient Domain-Independent Algorithm for Detecting Approximately Duplicate Database Records," Proc. Second ACM SIGMOD Workshop Research Issues in Data Mining and Knowledge Discovery (DMKD '97), pp. 23-29, 1997.

[7] Han, J., Kamber, M., "Data mining: concepts and techniques", 2 edn. Morgan Kaufmann (2006)

[8] Monge, A.E., "Matching algorithms within a duplicate detection system" ,IEEE Data Engineering Bulletin (2000)

[9] Naumann, F., Herschel, M., "An introduction to duplicate detection", Synthesis Lectures on Data Management, vol. 3. Morgan and Claypool Publishers (2010)

[10] P. Christen, "Development and User Experiences of an Open Source Data Cleaning, Deduplication and Record Linkage System," SIGKDD Explorations, vol. 11, no. 1, p. 3948, 2009.

[11] Brendan J. Frey, Delbert Dueck, "Clustering by passing messages between data points", Science, vol 315, 2007.