Efficient techniques for record comparision and de-duplication using Febrl framework

K.MALA,
M.E. (CSE),
srinivasan engg college,
perambalur

S.CHINNADURAI,
A.P,M.E(CSE),
srinivasan engineering college,
perambalur.

Abstract – **Record linkage is the problem of identifying similar records across different data sources. The similarity between two records is defined based on domain-specific similarity functions over several attributes. De-duplicating one data set or linking several data sets are increasingly important tasks in the data preparation steps of many data mining projects. The aim is to match all records relating to the same entity. Different measures have been used to characterize the quality and complexity of data linkage algorithms,and several new metrics have been proposed. An overview of the issues involved in measuring data linkage and de-duplication quality and complexity. A matching tree is used to overcome communication overhead and give matching decision as obtained using the conventional linkage technique. Developed new indexing techniques for scalable record linkage and de-duplication techniques into the febrl framework, as well as the investigation of learning techniques for efficient and accurate indexing.**

Keywords: ***data cleaning; similarity matching; record linkage; data mining pre-processing; febrl.***

1.INTRODUCTION

The most recent have observe a marvelous increase in the use of computerized databases for supporting a variety of company decisions. The data needed to support these decisions are often spread in diverse dispersed databases. In such cases, it may be necessary to linkage records in multiple databases so that one can merge and use the data pertaining to the same real world entity.

If the databases use the same set of design standards, this linking can easily be done using the primary Key, however, since these heterogeneous databases are usually designed and managed by different organizations, there may be no common candidate key for linking the records. Although it may be possible to use common non key attributes (such as name, address, and date of birth) for this purpose, the result obtained using these attributes may not always be accurate.

2. DATA CLEANING AND RECORD LINKAGE PROCESS

A general schematic outline of the record linkage process is given As most real-world data collections contain noisy, incomplete and incorrectly formatted information, data cleaning and standardization are important pre-processing steps for successful record linkage, and also before data can be loaded into data warehouses or used for further analysis or data mining. A lack of good quality data can be one of the biggest obstacles to successful record linkage and de-duplication. The main task of data cleaning and standardization is the conversion of the raw input data into well defined, consistent forms, as well as the resolution of inconsistencies in the way information is represented and encoded.

3. FEBRL FRAMEWORK

Python is an ideal platform for rapid prototype development as it provides data structures such as sets, lists and dictionaries (associative arrays) that allow efficient handling of very large data sets, and includes many

modules offering a large variety of functionalities. For example, it has excellent built-in string handling capabilities, and the large number of extension modules facilitate, for example, database access and graphical user interface (GUI) development. For the Febrl user interface, the PyGTK4 library and the Glade5 toolkit were use, which, combined, allow rapid platform independent GUI development

.

3.1) Input Data Initialization

In a first step, a user has to select if she or he wishes to conduct a project for (a) cleaning and standardization of a data set, (b) de-duplication of a data set, or (c) linkage of two data sets. The 'Data' page of the Febrl GUI will change accordingly and either show one or two data set selection areas. Several texts based data set types are currently supported, including the most commonly used comma separated values (CSV) file format. SQL database access will be added in the near future. Various settings can be selected, such as if a data set file contains a header line with field names (if not these field names can be entered manually); if one of the fields contains unique record identifiers; a list of missing values can be given (like 'missing' or 'n/a') that automatically will be removed when the data is loaded; and there are data type specific parameters to be set as well (such as the delimiter for CSV data sets).

3.2) Data Exploration

The 'Explore' page allows the user to analyses the selected input data set(s) in order to get a better understanding of the content and quality of the data to be used for a standardization, de-duplication or linkage project. In order to speed up exploration of large data sets, it is possible to select a sampling rate as percentage of the number of records in a data set.

3.3) Data Cleaning and Standardisation

The cleaning and standardization of a data set using the Febrl GUI is currently done separately from a linkage or de-duplication project, rather than as a first step. A data set can be cleaned and standardized and is written into a new data set, which in turn can then be de-duplicated or used for a linkage. When a user selects the 'Standardization' project type, and has initialized a data set on the 'Data' page, she or he can define one or more component standardizes on the 'Standardize'page.Currently, component standardizes are available in Febrl for names, addresses, dates, and telephone numbers. The name standardize uses a rule-based approach for simple names (such as those made of one given- and one surname only) in combination with a probabilistic hidden Markov model (HMM) approach for more complex names (Churches et al. 2002), while address standardization is fully based on a HMM approach (Christen and Belacic 2005). These HMMs currently have to be trained outside of the Febrl GUI, using separate Febrl modules. Dates are standardized using a list of format strings that provide the expected formats of the dates likely to be found in the un cleaned input data set. Telephone numbers are also standardized using a rules based approach. Each standardize requires one or several input fields from the input data set (shown on the left side of a standardize in the GUI), and cleans and segments a component into a number of output fields (three for dates, five for phone numbers, six for names, and 27 for addresses), shown on the right side in the GUI.

4. INDEXING DEFINITION

'QGramIndex', which uses sub-strings of length q (for example bigrams, where q = 2) to allow fuzzy blocking (Baxter et al. 2003); 'Canopy Index', which employs overlapping canopy clustering using TF-IDF or Jaccard similarity (Cohen and Richman 2002); 'String Map Index', which maps the index key values into a multi-dimensional space and performs canopy clustering on these multi-dimensional objects (Jin et al. 2003); and 'Suffix Array Index', which generates all suffixes of the index key values and inserts them into a sorted array to enable efficient access to the index key values and generation of the corresponding blocks (Aizawa and Oyama 2005).

For deduplication using 'BlockingIndex', 'Sorting Index' or 'QGram Index', the indexing step can be performed in an overlapping fashion with the field comparison step, by building an inverted index data structure while records are read from the input data set and their blocking key values are extracted and inserted into the index. The current record is compared with all previously read and indexed records having the same blocking key value. This approach can be selected by the user by ticking the 'De-duplication' indexing box. For a linkage, and using one of the three indexing methods mentioned above, the Big Match (Yancey 2002) approach can be selected, where first the smaller input data set is loaded and the

inverted index data structures are built in main memory, including all record attribute values required in the comparison step. Each record of the larger input data set is then read, its blocking key values are extracted, and all records in the same block from the smaller data set are retrieved from the index data structure and compared with the current record. This approach performs only one single pass over the large data set and does not require indexing, sorting or storing of any of its records. The user can tick the corresponding 'Big Match' indexing box when conducting a linkage project.

5. FIELD COMPARISON FUNCTIONS

The comparison functions to be used to compare the field values of record pairs can be selected and setup on the 'Comparison'. Each field comparison requires the user to select one of the many available comparison functions as well as the two record fields that will be compared. While one normally would select fields with the same content from the two data sets (for example, to compare suburb names with suburb names), it is feasible to select different fields (for example to accommodate for swapped given- and surname values).

6. WEIGHT VECTOR CLASSIFICATION

The last major step required is the selection of the method used for weight vector classification and setting of its parameters. Currently, Febrl offers six different classification techniques. The simple 'Fellegi Sunter' classifier allows manual setting of two thresholds . With this classifier, the similarity weights of the weight vector of each compared record pair are summed into one matching weight, and record pairs that have a summed weight above the upper classification threshold are classified as matches, pairs with a matching weight below the lower threshold are classified as non-matches, and those record pairs that have a matching weight between the two classification thresholds are classified as possible matches.

With the 'Optimal Threshold' classifier it is assumed that the true match status for all compared record pairs is known (i.e. supervised classification), and thus an optimal threshold can be calculated based on the corresponding summed weight vectors. The match status is assumed to have been generated by an exact comparison of one of the fields in the data set(s). For example, if a field 'entity id' contains the entity identifiers, an exact match of two records that refer to the same entity will result in a similarity value 1, while all comparisons of records that refer to two different entities will result in a similarity value of 0. Thus these similarity values can be used to determine the true match and non-match status of all weight vectors, which in turn can be used to find one optimal classification threshold (i.e. no record pairs will be classified as possible matches). Both the 'KMeans' and 'FarthestFirst' classifiers are based on unsupervised clustering approaches, and group the weight vectors into a match and a non-match cluster. Several methods for centroid initialization and different distance measures can be selected. It is also possible to use only a fraction of the weight vectors when calculating the clusters (using sampling), which will be useful when de-duplicating or linking large data sets that have resulted in a large number of weight vectors. These two classifiers also allow the selection of a 'fuzzy region', as described in (Gu and Baxter 2006), which will classify the weight vectors, and thus the corresponding record pairs, in the area half-way between the match and non-match centroids as possible matches. The Febrl user interface with the 'KMeans' classifier.

The 'SuppVecMachine' classifier uses a supervised support vector machine (SVM) and thus requires the user to provide the true match status (as described above for the 'Optimal Threshold' classifier) of weight vectors in order to be able to train this classifier. It is based on the lib svm library, and the most important parameters of the SVM classifier can be set in the Febrl GUI. Finally, the 'Two Step' classifier is an unsupervised approach which in a first step selects weight vectors from the compared record pairs that with high likelihood correspond to true matches and true non-matches, and in a second step uses these vectors as training examples for a binary classifier (Christen 2007). Several methods are implemented on how to select the training examples in the first step, and for the second step a SVM classifier or k-means clustering can be used. Experimental results have shown that this unsupervised approach to weight vector classification can a achieve linkage quality almost as good as fully supervised classification (Christen 2007).

7. PERFORMANCE AND RESULT

The different methods were executed several times with different partition/window sizes. To

obtain comparable results, the partition sizes for blocking were selected in such a way that there was always a corresponding window size with nearly the same total number of comparisons. This was achieved by sorting the tuples and cutting them in fixed size partitions. Additionally, an exhaustive comparison of all tuples was processed without any partitioning. This is especially interesting to see the impact of partitioning methods on the recall.

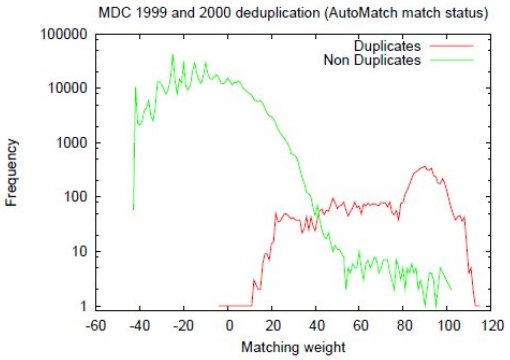

Fig 1: The density plot of the matching weights for a real-world administrative health data set. This plot is based on record pair comparison weights in a blocked comparison space. The smallest weight is -43, the highest 115. Note that the vertical axis with frequency counts is on a log scale

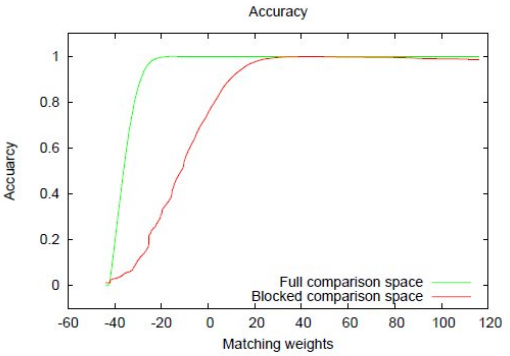

Fig 2a: Accuracy and matching weight

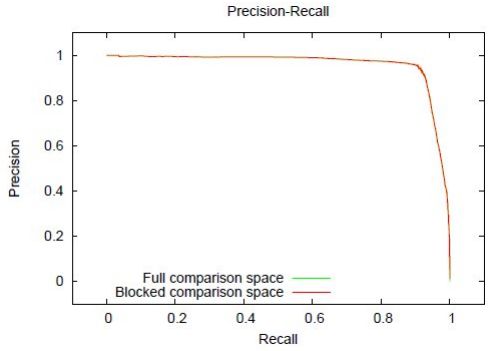

Fig 2b: precision and recall

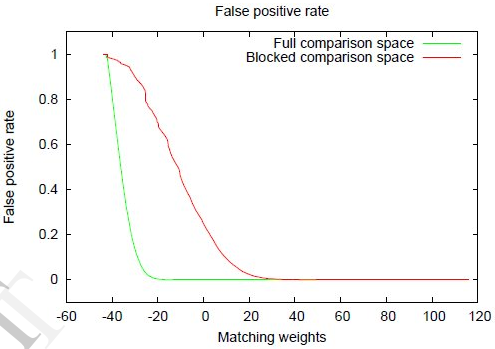

Fig 2c: matching weight and False Positive

Fig. 2. Quality measurements of a real world administrative health data set. The full comparison space (30; 698; 719; 310 record pairs) was simulated by assuming that the record pairs removed by blocking were normally distributed with matching weights between -43 and -10. Note that the precision-recall graph does not change at all, a ndthe F-measure graphs does change only slight. Accuracy and specificity are almost the same as both are dominated by the large number of true negatives. The ROC curve is the least illustrative graphs, which is again due to the large number of true negatives.

8. CONCLUSION

Febrl is an training tool suitable for new record linkage users and practitioners, and to conduct small to medium sized experimental linkages and de-duplications with up to several hundred thousand records. Within the health sector, it can be used alongside commercial linkage systems for comparative linkage studies; and for both new and experienced record linkage practitioners to learn about the many advanced

linkage techniques that have been developed in recent years and that are implemented in Febrl.

9. REFERENCES

[1]Baxter.R, Christen.P, and Churches.T, "A Comparison of Fast Blocking Methods for Record Linkage," Proc. ACM Workshop Data Cleaning, Record Linkage and Object Consolidation (SIGKDD '03), pp. 25-27, 2003

.[2] Bilenko.M, Basu.S, and Sahami.M, "Adaptive Product Normalization: Using Online Learning for Record Linkage in Comparison Shopping," Proc. IEEE Int'l Conf. Data Mining (ICDM '05), pp. 58-65, 2005.

[3] Bilenko.M and Mooney.R.J, "On Evaluation and Training-Set Construction for Duplicate Detection," Proc. Workshop Data Cleaning, Record Linkage and Object Consolidation (SIGKDD '03), pp. 7-12, 2003.

[4] Bilenko.M, .Kamath.B, and Mooney.R.J, "Adaptive Blocking: Learning to Scale up Record Linkage," Proc. Sixth Int'l Conf. Data Mining (ICDM '06), pp. 87-96, 2006.

[5] Clark.D.E, "Practical Introduction to Record Linkage for Injury Research," Injury Prevention, vol. 10, pp. 186-191, 2004.

[6]Churches.T, Christen.P, .K Lim, and Zhu.J.X, "Preparation of Name and Address Data for Record Linkage Using Hidden Markov Models," Biomed Central Medical Informatics and Decision Making, vol. 2, no. 9, 2002.

[7]Christen.P and Goiser.K, "Quality and Complexity Measures for Data Linkage and Deduplication," Quality Measures in Data Mining, ser. Studies in Computational Intelligence, Guillet.F and Hamilton.H, eds., vol. 43, Springer, pp. 127-151, 2007.

[8]Christen.P, "Febrl: An Open Source Data Cleaning, Deduplication and Record Linkage System With a Graphical User Interface," Proc. 14th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '08), pp. 1065-1068, 2008.

[9]Christen.P, "Automatic Record Linkage Using Seeded Nearest Neighbour and Support Vector Machine Classification," Proc. 14th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '08), pp. 151-159, 2008.

[10]Peter christen,"A Survey of indexing Techniques for Scalable Record Linkage And Deduplication", IEEE Transcations on Knowledge And Data Engineering , vol 24,no.9.september 2012.