MEDINFO 2010
C. Safran et al. (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-588-4-1127

Record Linkage System in a Complex Relational Database - MINPHIS Example

Philip Achimugu[a], Abimbola Soriyan[b], Oluwatolani Oluwagbemi[a], Anu Ajayi[b]

[a]*Department of Computer Science, Lead City University, Ibadan*
[b]*Department of Computer Science and Engineering, Obafemi Awolowo, Ile-Ife*

Abstract

In the health sector, record linkage is of paramount importance as clinical data can be distributed across different data repositories leading to duplication. Record Linkage is the process of tracking duplicate records that actually refers to the same entity. This paper proposes a fast and efficient method for duplicates detection within the healthcare domain. The first step is to standardize the data in the database using SQL. The second is to match similar pair records, and third step is to organize records into match and non-match status. The system was developed in Unified Modeling Language and Java. In the batch analysis of 31, 177 "supposedly" distinct identities, our method isolates 25, 117 true unique records and 6, 060 suspected duplicates using a healthcare system called MINPHIS (Made in Nigeria Primary Healthcare Information System) as the test bed.

Keywords:

Record linkage, Data mining, Duplicates, Databases, MINPHIS

Introduction

Many private and public organizations in the health sector capture, store, process and analyze fast-growing amounts of data with millions of records. The records are made up of patient's bio-data and health records. Linking and aggregating records that relate to the same person from several databases is becoming increasingly important as information from multiple sources needs to be integrated, combined or linked in order to allow detailed data analysis or mining or warehousing. The aim of such linkages is to match all records relating to the same entity for better informed decisions at various levels.

The basic methods compares name and address information across pairs of files to determine those pair of records that are associated with the same entity. The most sophisticated methods use information from multiple lists [7]; create new functional relationships between variables in two files that can be associated with new metrics for identifying corresponding entities [8] or use graph theoretic ideas for representing linkage relationships as conditional random fields that can be partitioned into clusters representing individual entities [4].

The main challenge in this task is designing a function that can resolve when a pair of records refers to the same entity in spite of various data inconsistencies. Data quality has many dimensions or qualities, one of which is accuracy. Accuracy is usually compromised by errors accidentally or intentionally introduced in a database system. These errors result in inconsistent, incomplete or erroneous data elements. In order to improve the accuracy of the data stored in a database system, we need to compare them either with their real world counterparts or with other data stored in the same or a different system.

Materials and Methods

This section describes the material and method that were used in achieving the desired or set goal.

MINPHIS is an acronym that stands for Made in Nigeria Primary Healthcare Information Systems; a software system that was collaboratively developed by the Health Information Systems Research and Development Unit of the Obafemi Awolowo University Ile-Ife, Nigeria and the Health Information Systems Research and Development Unit of the University of Kuopio Finland in 1989. Currently, MINPHIS has been deployed to over eleven (11) teaching and specialist hospitals in Nigeria. Over 30, 000 records were pulled out of MINPHIS database deployed at the Obafemi Awolowo University Teaching Hospitals Complex for testing the system developed.

Given databases A and B, record linkage finds or detects the common entity between them, (figure 1). Each record from A potentially has to be compared with all the records from B. The total number of potential record pair comparisons thus equal to the product of the size of the two databases.

Sur-name	Other Names	Sex	DOB	UID
Peg	R. Smith	M	24-09-1964	01-001
Joe	Smith	M	26-10-1983	01-008
John	Thereza	F	26-06-1984	01-007

Database A

Matches

Sur-name	Other Names	Sex	DOB	UID
Robert	P. Smith	M	24-09-1964	01-016
Joe	Smithy	M	26-10-1983	01-008
Jon	Theresa	F	26-06-1984	01-007

Database B

Figure 1- Record Linkage Example

To reduce the large amount of potential record pair comparisons, our system employs a technique called 'blocking'; a single record attribute or a combination of attributes called blocking key or variable was used to split the database into blocks. Therefore, we sorted the records in our database alphabetically using surname, sex and date of birth, so that only records that falls within the same block are compared with their counterparts.

All records having the same value in the blocking key were inserted into one block and candidate record pairs are generated only from records within the same block. While the aim of blocking is to reduce the number of record pair comparisons made as much as possible by eliminating pairs of records that obviously are not matches, it is also important that no true matches are removed by the blocking process. That was why; we had to block records alphabetically to allow scalability or robustness of the blocking process and also to ensure that no true match is missed.

Two main issues are considered when blocking key is defined:

1. The error characteristics of the attributes used in blocking keys will influence the quality of the generated candidate record pairs. Therefore, attributes containing the fewest errors or missing values should be chosen as any error in an attribute value in a blocking key will potentially result in a record being inserted into the wrong block, thus missing true matches.
2. The frequency distribution of the values in the attributes used as blocking key will affect the size of block generated. So, if *m* records are in a block from database A and *n* records in the same block from database B, then *m* x *n* record pairs will be generated from this block. The largest block will dominate the execution time of the comparison step as they will contribute very large numbers of record pairs.

In order to address the problems enumerated above, we developed a string matching function that is embedded in the record linkage system algorithm to cater for strings with typographical errors as a result of keystroke mistakes or fatigue during the data entry process. This will enhance the blocking process because true matches will not fall into wrong block. The string matching function compares two strings say **JULIUS Babatunde** and **JULIUUS Babatunde; or Achimugu Philip** and **Chimugu Philip,** and calculates the number of common character and transposition. So if the total number of common characters between the two strings is more than three quarters of the length of the shorter string, then the function suspects and reports a likelihood of typographical error in the two strings, before other attributes such as Date of Birth, Sex and address information are finally compared to determine the status quo of such entities.

In this experiment, the blocking technique used for our health database allows the size of blocks to be controlled directly through parameters. All the candidate record pairs generated by the blocking process are compared by the comparison function applied to one or more (or a combination of) record attributes.

Each comparison returns a numerical similarity value called 'matching weight' (1 if the strings are similar or agreeing and 0 if the strings are not similar or disagreeing). A vector is formed for each compared record pair containing all the values calculated by the comparison function. These vectors are then used to calculate record pairs into match, non-match and possible match based on the decision model developed. Therefore, record pairs that were removed by the blocking process are classified as non-match or unique records without being compared explicitly. Figure 2 depicts the record linkage process.

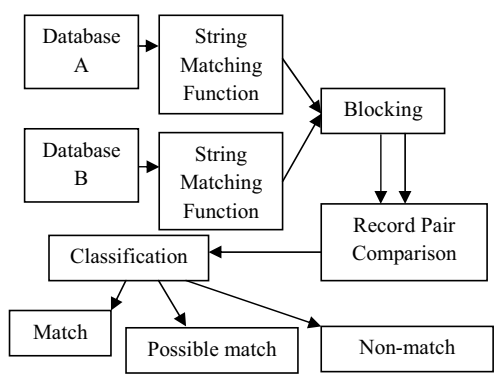

Figure 2- The Record Linkage Process

Experimental Results

To evaluate the performance of this algorithm, normalized measures such as precision and recall were used to determine the efficiency of the algorithm. The experiments performed consisted of finding matches between two data sets.

Test Data

The test used a simplified dataset containing all bio-data of patients from the MINPHIS database. The datasets contain a range of identifying information such as names, address, diagnosis, referrals, ward allotment etc for 31, 177 supposedly distinct patients.

For the purpose of this research however, the attributes that were extracted from the MINPHIS database for experimental evaluation are Hospital Number, Surname, firstNames, Sex, Date of Birth and Address information. Records representing the same entity based on the static decision rules were given the same value of 1 while those that are not were given the value 0. That was the criteria for duplicate detection.

We therefore, conclude that two records are match if they correspond in names (Surname, FirstNames), date of birth, sex and address.

Evaluation of the System

The evaluation of the algorithm encompasses two main issues: (1) the accuracy and (2) the behaviour of the similarity threshold k. First, the distances between all possible pairs of records (r_i, r_j) are computed and stored in a matrix. Then, for each value of k, the total number of true positives (TP), false positives (FP), false negatives (FN) and true negatives (TN) matches are computed using the formula below:

$$\text{Recall} = \frac{TP}{TP + FN} \quad \textbf{(1)}$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad \textbf{(2)}$$

Reference Tables

We used eight blocks for the experiment. The first reference table is the combination of unique given names and surnames from the dataset giving a reference table size of 3415 unique names. The second reference table is every second name from the first reference table, starting with the first name, while the first reference table is every second name starting with the second name. The fourth reference table contains unrelated data, the unique surnames and given names from the MINPHIS database.

Table 1- Result of the System

Block	Correct Matches	Correctly Unlinked	Incorrect Matches	Precision	Recall
1	3415	302	400	0.92	0.90
2	3308	292	295	0.92	0.91
3	3936	200	225	0.95	0.94
4	3501	109	305	0.96	0.91
5	3566	190	300	0.95	0.92
6	3486	217	320	0.94	0.91
7	3299	130	178	0.96	0.95
8	3175	17	11	1.0	1.0

Analysis of the Result

As expected there is a trade-off between Precision and Recall when the threshold k is varied. Its optimum value is considered to be in the intersection of the two curves (Figure 3). The system employs a method called **blocking** in order to reduce large comparisons between potential duplicate records, so only records that falls within the same block or neighbourhood are compared and tracked for duplicate detection hence, decrease in computational time. From table 1 therefore, it is deduced that the system produced a high level of quality duplicates detection as evidenced in the values for precision and recall. In Block one, it is observed that only 400 incorrect matches was found after retrieving 3415 records, that is, matches that could not be tracked by the static decision rules embedded in the algorithm and the final match status is determined by the human expert. It goes on through all the rest of the blocks until the information retrieval process is completed. Although, the system is automated but the final decision for records that falls under the possible match category is determined by the human expert. This is important because patients in health organizations are seen as owners of their medical records; therefore, adequate care must be taken to ensure that data are not altered in any way throughout the record linkage process.

Furthermore, Figure 3 depicts the graphical representation between precision and recall. It shows a significant increase in precision as regards the quality of duplicates detected by the enhanced system. For example, it is approximately 0.18, corresponding to 0.95 of precision and recall (Figure 3). In comparison with exact record matching, which is equivalent to the case k = 0, the approximate record matching (with higher values for k) provided a good gain in Recall, without significant loss in Precision. But, when the dataset from MINPHIS database was tested on the algorithm which has embedded a string matching function that caters for typographical errors in candidate's names, exact record matching obtained a Recall of only 40% (Figure 4).

There are also large regions (0.14 < k < 0.19, for the

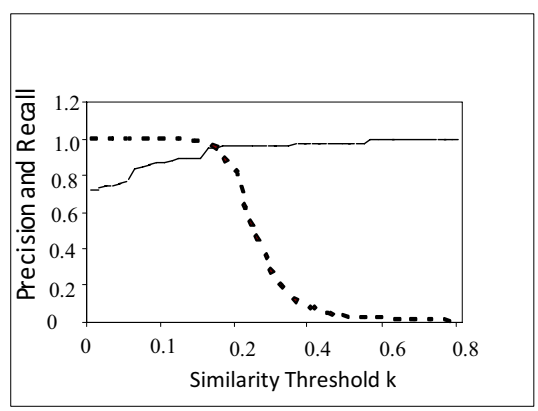

Figure 3- Tradeoff between Precision and Recall

algorithm; and $0.09 < k < 0.19$, for the tradeoff between precision and recall), where both precision and recall are high (greater than 0.9). For a while, this allows some freedom and safety in the choice of k.

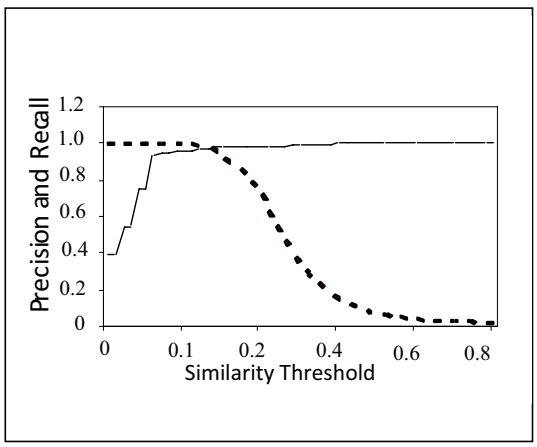

Figure 4- MINPHIS Dataset Bahaviour in the Record Linkage Algorithm

Discussion

The initial idea for record linkage was conceived by Halbert Dunn in 1946 who was the then chief of U.S. National Office of Vital Statistics. He used the term to refer to linking vital records, such as birth and death certificates, pertaining to a single individual [2]. Computerized record linkage was proposed a decade later when Howard Newcombe and colleagues used computers to link vital records in an effort to track hereditary diseases. The theory of record linkage was further expanded by Ivan Fellegi and Alan Sunter who demonstrated that probabilistic decision rules were optimal when the comparison attributes are conditionally independent [6]. Our method gives those specialists responsible for merging similar records a representative view to show them how close records in some homogenous or heterogeneous sets are. Additionally, the algorithm and underlying database support real-time detection of duplicate records. This can help to avoid the creation of duplicate records by alerting the user that several neighbour records already exist. This real-time use could also be used in multi criteria searches for identities and a simple as well as easy to use front end algorithm was employed in the implementation of the record linkage system so that short response times are achieved. Response time is closely related to optimization of the algorithm and especially the blocking part. Its improvement allows the reduction in the number of potential duplicates to be tested by the main algorithm.

Conclusion

In this paper, a record linkage system for health information systems was developed and applied to health informatics in developing countries (particularly Nigeria). The methodology implored for achieving our goal is discussed herein and we believe that the result would be useful and the system more efficient than existing ones.

References

[1] Baxter R, Christen P, and Churches T. A Comparison of Fast Blocking Methods for Record Linkage. Proceedings of the ACM Workshop on Data Cleaning, Record Linkage and Object Identification, Washington, DC, August 2003.

[2] Dunn HL. Record Linkage, American Journal of Public Health 1946: 36, 1412-1416.

[3] Fellegi I P. and Sunter A B. A Theory for Record Linkage. Journal of the American Statistical Association 1969: 64, 1183-1210.

[4] McCallum A, Nigam K, and Unger L H. Efficient Clustering of High-Dimensional Data Sets with Application to Reference Matching, in Sixth ACM Conference of Knowledge Discovery and Data Mining 2000: 169-178.

[5] McCallum A, and Wellner B. Object Consolidation by Graph Partitioning with a Conditionally-Trained Distance Metric. Proceedings of the ACM Workshop on Data Cleaning, Record Linkage and Object Identification, Washington DC, August 2003.

[6] Newcombe HB, Kennedy JM, Axford SJ, and James AP, "Automatic Linkage of Vital Records. Science 1959: 130, 954-959.

[7] Scheuren F and Winkler WE. Regression Analysis of Data Files that are Computer matched, II, Survey Methodology 1997, 23, 157-165, http://www.fcsm.gov/working-papers/scheuren_part2.pdf.

[8] Winkler WE. The State of Record Linkage and Current Research Problems. Statistical Society of Canada, Proceedings of the Survey Methods Section 1999a, 73-80 (longer version also available at http://www.census.gov/srd/www/byyear.html).

[9] Winkler WE. Issues with Linking Files and Performing Analyses on the Merged Files. American Statistical Association, Proceedings of the Sections on Government Statistics and Social Statistics 1999, 262-265

Address for Correspondence

Achimugu Philip
Department of Computer Science, Room 118, Lead City University, Ibadan. E-Mail: check4philo@yahoo.com. Mobile Phone: +23480 5289 1845.