The New Navigators: from Professionals to Patients
R. Baud et al. (Eds.)
IOS Press, 2003

A Methodology for Multi-source Record Linkage

E.M. Kerkri [a,b], C. Quantin [a], T. Grison [b] and K. Yétongnon [b]

[a] *Dijon University Hospital, Medical informatics Department, Dijon, France*
[b] *University of Burgundy, Electronic and Informatic Engineering Laboratory, Dijon, France*

Abstract

As patient's medical data is disseminated in different health structures, building a data-warehousing system at a regional level, for healthcare co-operation, has some specific requirements compared to intra healthcare structure data-warehousing projects. In particular, there is no common patient electronic record's identifier among all the sources. Hence, data integration and Record Linkage is the most important issue in such projects.

In this paper, we present Background on Information integration with the two main approaches of the problem: Schema Integration approach and Entity reconciliation approach.. We show the limits of each of the two approaches and propose an approach based on theory of a set partitioning and instance identification.

Keywords: *Record Linkage; Instance identification; Entity reconciliation; Partition.*

1 Introduction

Co-operation in healthcare area concerns many partners: hospital schools, anti-cancer centers, laboratories of biology, radiology centers, doctors' offices; anatomy pathologic laboratories, public and private hospitals. These partners often use different codification and classifications systems to encode their data. Thus, patient's medical data is disseminated in different health structures.

Building a data-warehousing system at a regional level, for healthcare co-operation, has some specific requirements compared to intra healthcare structure data-warehousing projects [1,4]. In particular, there is no common patient electronic record's identifier among all the sources. Hence, Record Linkage is the most important issue in such projects.

The aim of file linkage[5,6] is to gather all information coming from different sources and concerning the same patient. Two types of linkage errors are of concern : erroneous links of notifications from two distinct patients, also called homonym errors, and failure to link multiple notifications on the same patient, also called synonym errors [5].

To reduce the impact of typing errors, a spelling treatment [5] has been introduced in the anonymity process, before the hash coding. The principle is to transform the spelling of names according to phonetic rules.

Moreover, the linkage takes into account several identification variables such as, for example, first and last names, date of birth, gender and zip code. However, some variables provide more information and more reliability than others. As a consequence, Jaro [5,6] proposed to associate a weight to each variable according to the reliability of the information provided by this variable. The weight given to date of birth is then higher than that given to gender, as two identical dates of birth are more discriminant than two identical

genders. The more reliable a variable, the greater the weight will be, as the weight is computed from the logarithm of the ratio between the sensitivity and one minus the specificity of the studied variable (likelihood ratio). A composite weight is obtained by summing the weight of each variable.
The aim of the linking of two files F1 and F2 is to classify each pair of records obtained from crossing F1xF2 as belonging to one of two sets: the set of matched record pairs M, and the set of unmatched record pairs U. From a statistical point of view, this problem is equivalent to the analysis of a finite mixture of two sets M and U, in proportions p and (1-p). The belonging of a pair to M (resp. U) is supposed to follow a binomial distribution characterized by the parameter m (resp. u). The m probability can be defined as the probability of agreement of the two records of the pair, for the considered variable, knowing that the two records correspond to the same individual. The estimation of the parameters (m, u, p) is obtained through the maximization of the data likelihood. After having ordered the 2n possible configurations of agreement and disagreement of pairs of records composed of n variables by the composite weight, one can compute the cumulative distribution functions of these configurations, conditionally to belonging to the sets M and U. Two threshold values can then be computed from which three sets of possible decisions are determined as follows: the pair is a match ; no determination is made ; the pair is not a match.
As we can see, this approach consider only two sets of records at one time. In Multi-source Co-operation, we deal with several sets of records in the same time. In this paper, we present a background on Information integration with the two main approaches of the problem: Schema Integration approach and Entity reconciliation approach.. We show the limits of each of the two approaches and propose an approach based on theory of a set partitioning and instance identification.

2 Background on Information integration

Record Linkage problem could be resolved in two main approaches. Firstly, schema integration uses schema level information and structural conflicts to resolve the semantic heterogeneity problem so that the target databases could be populated with data. Secondly, entity reconciliation uses instance level information to merge and integrate data into the target database.

2.1 Schema Integration approach

The schema integration approach uses schema level information and structural conflicts to resolve the semantic heterogeneity problem. It combines different user views into a single global view. Three popular approaches to databases integration are (i) The global schema approach produces a single logical view of the integrated databases; (ii) Federated schema approach integrates multiple export-schemas from each local database. It is based on the federated database approach; (iii) Semi-decentralized approach integrates both global and federated schema approaches.
Batini and Lenzerini [7] discuss a methodology for schema integration based on entity-relationship model. They provide a survey of schema integration literature and use a four-phase integration process: (i) preintegration, (ii) comparison of the schemas, (iii) conformation of the schemas, and (iv) merging and restructuring of the schemas. Larson et al. [8] develop a theoretical framework for establishing equivalence between attributes for the purpose of schema integration. Gotthard et al. [9] discuss a system-guided view integration process that suggests similarities between the structures of two schemas and ask a human integrator to either accept or reject that structures are the same.
The schema integration solution enables resolving structural heterogeneity. well adapted to dynamic access to data among heterogeneous sources.

2.2 Entity reconciliation approach

Given N sets of records to integrate. At the instance level, two main problems may occur:

- Entity identification problem which is the consequence of the missing of a common key (or identifier) among different sources,
- Data incoherency due to the name or address changes (by example) and / or existing of keyboard errors.

Resolving these two errors is the main issue for heterogeneous databases integration at instance level. We have to avoid either gathering instances not corresponding to the same real world objects or separating instances that not have to be so. This problem could be considered in different point of view, mainly the entity identification and reconciliation approach and data clustering approach.

The entity identification problem is to match object instances from different databases, which correspond to the same real world entity.

Lim et al. [10] present a 2-step entity identification process in which attributes for matching tuples may be missing in certain tuples, and then need to be derived prior to the matching. To match tuples, they require identity rules that specify the conditions to be satisfied by a pair of tuples, from different databases, before they can be considered as modeling the same real-world entity. Tejada et al. [11] propose a solution based on learning object identification rules for information integration. They have developed an object identification system, which compares the object shared attributes in order to identify matching objects. The system learns to tailor mapping rules, through limited user input, to a specific application domain.

Phonetic matching is used to identify strings that are likely to have similar pronunciation, regardless of their spelling. Liam et al. [12] propose a technique for phonetic matching approach based on information derived from a pronunciation dictionary.

To identify approximately duplicate records in databases, Bilenko et al. [13] propose a domain-independent method using machine learning. First, trainable distance metrics are learned for each field. Second, a classifier is employed that uses several diverse metrics for each field as distance features and classifies pairs of records as duplicates or non-duplicates.

To achieve the entity matching in heterogeneity heterogeneous databases, Dey et al.[14] propose a decision theoretic model that uses a distance-based measure to express the similarity between two entity instances.

To identifying equivalent data instances in federated databases context, Si et al. [15] employ a probabilistic model, which utilizes historical database update information to estimate the degree of similarity between candidate data instances from different database components. They employ transaction history (log) information to this end, which is typically already available in the component database systems.

3 Multisources Record Linkage: a partition based approach

Given N sets of records to integrate. At the instance level, two main problems may occur. The first one is Entity identification problem that is the consequence of the missing of a common key (or identifier) among different sources. The second concerns the data incoherency, which is due to the name or address changes (by example) and / or existing of keyboard errors.

Resolving these two errors is the main issue for heterogeneous databases integration at instance level. We have to avoid either gathering instances not corresponding to the same real world objects or separating instances that not have to be so.

3.1 Principle of the method

Let A_1, A_2, .., A_N be N data sets issued from heterogeneous sources. Let f_1, f_2, .., f_K, be K attribute enabling object characterization. This set of fields is not necessarily unique. So that other combinations of instance fields may be used for this target. This set of fields must exist in each information sources. If not so, we need to work on sub-sets and to process the Record Linkage in some steps.

To perform Record Linkage among these N data sets, we may use three ways: incremental recursive merging, parallel recursive merging and global merging

Incremental recursive merging consists in integrating data from A_1 and A_2, the result is integrated with A_3 and so on. In this case, when performing the linkage between two sets, we do not take into account the other sets. The consequence may be wrong instance linkage in some limit cases.

Parallel recursive merging consists in integrating concurrently data sets by pairs, if the result is one set, the program stops else a new step is executed. This method presents the same disadvantage as the first one, even if it is time cost less.

Global merging consists to consider a new set $E = \cup A_i$, $i = 1,N$ and to process a partition of the set E that consists to transform it into a union of subsets $E = \cup V_i$, $i = 1,P$, where each subset V_i contains instances representing the same real-world entitie. We present bellow the principle of this method.

Let D be a distance defined on E as follow:

$$D : E \times E \longrightarrow R^+$$

$$(m_1, m_2) \longrightarrow \sum w_j \, d_j(m_{1j}, m_{2j}) \; ; \; j = 1,K$$

w_j and d_j are respectively the weight and the distance of the j^{th} field (or variable).

The neighborhood of a given instance m is calculated

$$V(m) = \{p \in E \; / \; D(m,p) < \delta \}$$

δ is the threshold of acceptance of the belonging of p to V(m).

$$\delta = \sum w_j \, \delta_j \, max(d_j) \qquad j = 1,K$$

where δ_j is the threshold dedicated to the file number j adapted to its type and to the nature of the distance d_j and $max(d_j)$ is the maximum possible value for d_j.

As we can see, different problems are posed by this method. The choice of relevant attribute distance functions is may be the most complex because of the heterogeneity of the different attributes. We need also to compute the weights of each one of the attributes. The threshold calculation is the tird problem te resolve before performing the decision process. All this aspects are presented bellow

3.2 Choice of distances

The choice of relevant attribute distance functions is critical task for estimating the distance between two instances to decide if yes or not they represent the real-world object. The attribute distance function depend on the nature / type of the attribute possible values.

- Binary data value, the gender by example, the simplest case and the distance function is binary and defined by: if x=y then d(x,y)=0 else d(x,y)=1.
- For the discrete data value, the distance function may be discrete as well but the values depend on the data domain. For example, the Zip Code (in France) is a five-figure number (zc). The two left numerals (zcl) indicate the department code when the three right ones (zcr) indicate the commune code in the considered department. The distance function could be defined as: (if (zcl1<>zcl2) then d(zc1,zc2)=2 else-if (zcr<>zcr2) then d(zc1,zc2)=1 else d(zc1,zc2)=0)
- Distance functions for temporal attributes, such as the Date of Birth, could be defined in different manners. It could have binary values depending if the attribute values do

match or no. Or it could be more complex if we chouse to compare the date, the month and the year of birth together or not.

- For string values attributes such as the name or the address there many ways to compare two strings. The edit distance, the binary distance and the phonetic distance are some of possible distances to use. The edit distance is defined as the smallest number of insertions, deletions, and substitutions required to change one string into another. The phonetic distance consists to apply either edit distance or binary distance to the string after transforming it to a phonetically. Soundex uses codes based on the sound of each letter to translate a string into a canonical form of at most four characters, preserving the first letter. Phonix is a Soundex variant. Slightly different set of codes is used, and prior to mapping about 160 letter-group transformations are used to standardize the string e.g. from *x* to *ecs*. Editex is a phonetic distance measure that combines the properties of edit distances with the letter-grouping strategy used by Soundex and Phonix. Editex groups letters that can result in similar pronunciations, but doesn't require that the groups be disjoint and can thus reflect the correspondences between letters and possible similar pronunciation more accurately.

3.3 Weights of attributes

Computing the weights of the attributes requires more insight into the problem, to see how each attribute's scores should weigh for the overall score. In probabilistic methods [], weights are used to measure the contribution of each attribute to the probability of making a correct judgment for a pair of records matching. The weight value depend if the attribute agree, $w = \log_2 (m/u)$, or disagree, $w = \log_2 ((1-m)/(1-u))$ where m (respectively u) is the probability that the pair match (respectively do not match). In some cases one may use the relative density weighting function $w = N_d / N_t$ where N_d is the number of distinct values for the considered attribute and N_t is the set cardinality. If we want to privilege one are some attributes to accentuate their discrimination, we may give arbitrary weights to facilitate the set partition.

3.4 Threshold calculation

The threshold calculation is a critical stage for deciding Record Linkage. If the user is expert about the nature of the data he must be associated in the task. Otherwise, the error ratio accepted for the data could be used as indicator. In general, a statistical error ratio (5%) could be used. In some cases we need to perform exact matching and need to examine instances that cannot belong to a subset but it could be if the threshold was higher. In such case, a second threshold is needed and the result will contain the subsets with for each of them the list of limit instances.

3.5 Decision process

When all of the previous steps are performed the process as follows:

```
The attribute weights are calculated (if not given directly
  by the user)
REPEAT
The most density for the highest weighted attribute is
  localized (by frequency sort)
Choose an instance from this subset
Calculate the corresponding neighborhood Vi (and the border
  list of instances Bi)
E = E - Vi
UNTIL (E=∅) or card (Vi)=1
```

If $\cup B_i <> \varnothing$ then the concerned instances would have to be delaminated by the user or the DBA and a second pass will be processed. If $B_{i1} \cap B_{i2} <> \varnothing$ then a conflict is occurred for

the appurtenance of one or some instances to V_{i1} or V_{i2}. If the second pass cannot resolve the conflict. The user or the DBA will have to decide for the concerned instances.

4 Conclusion

An efficient cooperation between healthecare structures, at a regional level for patient follow-up and either medical or epidemiological studies, have to be based on the two important priclipales. The first priciple concern the data warehousing architecture that enables the integration treatment of medical data in a secure and safe way for the information sources. The second priciple concern the capability of a high quality medical record linkage. This two principles are included in this paper.
In the ongooing inplementation of the Multisource Recorde Linkage proposed in this paper aims to compare this method to probalistic and direct methods at two levels. It will examine in particular the quality and the time costs, which are the main indictors of the Recorl Linkage procedures for Multisource and Big sets of medical data. The results of such investigations will be presented in the MIE congress.

References:

[1] Kimball R. The Data Warehouse Toolkit, French version by Raimond C., International Thomson Publishing France, Paris, 1997.
[2] Sakaguchi T. and Frolik Mark N. A Review of the Data Warehousing Literature. http://www.people.memphis.edu/~tsakagch/dw-web.htm. Jan. 31, 1996.
[3] S. Chaudhuri and U. Dayal, An overview of Data Warehousing and OLAP Technology. SIGMOD Record, Vol. 26, No. 1, pp. 65-74, 1997
[4] Kerkri E.M. et al., An approach for Integrating Heterogeneous Information Sources in a Medical Data Warehouse, Journal of Medical Systems, Volume 25, N° 3, 2001 pp 167-176
[5] Jaro M.A., Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida, Journal of the American Statistical Association, 1989.
[6] Jaro M.A., Probabilistic-Linkage of large public health data files, Statistics in Medecine,1995 ; 14 : 491-498.
[7] Batini H., Lenzerini M., "*A Methodology for Data Schema Integration in the Entity-Relationship Model*", IEEE Transactions On Software Engineering, Vol. SE-10, n6, November 1984, 650-664
[8] Gotthard et al.: System-guided view integration for Object Oriented databases, IEEE Transactions on knowledge and data Engineering, Vol. 4, N°. 1, February 1992. pp 1-22
[9] Lim E.-P. et al., Entity identification in database integration. In International Conference on Data Engineering, pages 294--301, Los Alamitos, Ca., USA, April 1993. IEEE Computer Society Press.
[10] Larson, J., S. B. Navathe, and R. Elmasri. *A Theory of Attribute Equivalence in Databases with Application to Schema Integration*, IEEE Transactions on Software Engineering, Vol. 15, No. 4, April 1989.
[11] Tejada S.. Knobleock C.-A., Minton S.. Learning Object Identification Rules for Information Integration, Information Systems Vol. 26, No. 8, pp. 607-633, 2001
[12] Liam H.-W. and Jos L.., Using a Pronunciation Dictionary and Phonetic Rules for Name Matching Applications. http://goanna.cs.rmit.edu.au/~jz/sci/
[13] Bilenko M. and Mooney R.-J., Learning to Combine Trained Distance Metrics for Duplicate Detection in Databases, Technical Report AI 02-296, Artificial Intelligence Lab, University of Texas at Austin, February 2002
[14] Dey D. et al. A Distance-Based Approach to Entity Reconciliation in Heterogeneous Databases, IEEE Transactions on knowledge and data Engineering, Vol. 14, N°. 3, May/June 2002. pp. 567-582
[15] Antonio Si A. et al.,On Using Historical Update Information for Instance Identification in Federated Databases, First IFCIS International Conference on Cooperative Information Systems (CoopIS'96), June 19 - 21, 1996 Brussels, Belgium

Address for correspondence
El Mostafa KERKRI, Service d'Informatique Médicale
CHU - BP 1542 - 21034 Dijon Cedex
E-mail : emkerkri@u-bourgogne.fr