Febrl – A Parallel Open Source Data Linkage System

http://datamining.anu.edu.au/linkage.html

Peter Christen[1⋆], Tim Churches[2], and Markus Hegland[3]

[1] Department of Computer Science, Australian National University, Canberra ACT 0200, Australia, peter.christen@anu.edu.au
[2] Centre for Epidemiology and Research, New South Wales Department of Health, Locked Mail Bag 961, North Sydney NSW 2059, Australia, tchur@doh.health.nsw.gov.au
[3] Centre for Mathematics and its Applications, Mathematical Sciences Institute, Australian National University, Canberra ACT 0200, Australia, markus.hegland@anu.edu.au

Abstract. In many data mining projects information from multiple data sources needs to be integrated, combined or linked in order to allow more detailed analysis. The aim of such linkages is to merge all records relating to the same entity, such as a patient or a customer. Most of the time the linkage process is challenged by the lack of a common unique entity identifier, and thus becomes non-trivial. Linking todays large data collections becomes increasingly difficult using traditional linkage techniques. In this paper we present an innovating data linkage system called *Febrl*, which includes a new probabilistic approach for improved data cleaning and standardisation, innovative indexing methods, a parallelisation approach which is implemented transparently to the user, and a data set generator which allows the random creation of records containing names and addresses. Implemented as open source software, *Febrl* is an ideal experimental platform for new linkage algorithms and techniques.

Keywords: Record linkage, data matching, data cleaning and standardisation, parallel processing, data mining preprocessing.

1 Introduction

Data linkage can be used to improve data quality and integrity, to allow re-use of existing data sources for new studies, and to reduce costs and efforts in data acquisition for research studies. In the health sector, for example, linked data might contain information which is needed to improve health policies, information that is traditionally collected with time consuming and expensive survey methods. Linked data can also help in health surveillance systems to enrich data that is used for pattern detection in data mining systems. Businesses routinely

⋆ Corresponding author

H. Dai, R. Srikant, and C. Zhang (Eds.): PAKDD 2004, LNAI 3056, pp. 638–647, 2004.

deduplicate and link their data sets to compile mailing lists. Another application of current interest is the use of data linkage in crime and terror detection.

If a unique entity identifier or key is available in all the data sets to be linked, then the problem of linking at the entity level becomes trivial, a simple *join* operation in *SQL* or its equivalent in other data management systems is all that is required. However, in most cases no unique key is shared by all of the data sets, and more sophisticated linkage techniques need to be applied. These techniques can be broadly classified into *deterministic* or rules-based approaches (in which sets of often very complex rules are used to classify pairs of records as *links*, i.e. relating to the same person or entity, or as *non-links*), and *probabilistic* approaches (in which statistical models are used to classify record pairs). Probabilistic methods can be further divided into those based on *classical* probabilistic record linkage theory as developed by *Fellegi & Sunter* [6], and newer approaches using maximum entropy, clustering and other machine learning techniques [2,4, 5,10,12,14,19].

Computer-assisted data linkage goes back as far as the 1950s. At that time, most linkage projects were based on *ad hoc* heuristic methods. The basic ideas of probabilistic data linkage were introduced by *Newcombe & Kennedy* [15] in 1962 while the theoretical foundation was provided by *Fellegi & Sunter* [6] in 1969. The basic idea is to link records by comparing common attributes, which include person identifiers (like names, dates of birth, etc.) and demographic information. Pairs of records are classified as *links* if their common attributes predominantly agree, or as *non-links* if they predominantly disagree. If two data sets **A** and **B** are to be linked, record pairs are classified in a product space $\mathbf{A} \times \mathbf{B}$ into M, the set of true matches, and U, the set of true non-matches. *Fellegi & Sunter* [6] considered ratios of probabilities of the form

$$R = \frac{P(\gamma \ \varepsilon \ \Gamma | M)}{P(\gamma \ \varepsilon \ \Gamma | U)}$$

where γ is an arbitrary agreement pattern in a comparison space Γ. For example, Γ might consist of six patterns representing simple agreement or disagreement on (1) given name, (2) surname, (3) date of birth, (4) street address, (5) suburb and (6) postcode. Alternatively, some of the γ might additionally account for the relative frequency with which specific values occur. For example, a surname value *"Miller"* is normally much more common than a value *"Dijkstra"*, resulting in a smaller agreement value. The ratio R or any monotonically increasing function of it (such as its logarithm) is referred to as a *matching weight*. A decision rule is then given by

if $R > t_{upper}$, then	designate a record pair as *link*
if $t_{lower} \leq R \leq t_{upper}$, then	designate a record pair as *possible link*
if $R < t_{lower}$, then	designate a record pair as *non-link*

The thresholds t_{lower} and t_{upper} are determined by a-priori error bounds on false links and false non-links. If $\gamma \ \varepsilon \ \Gamma$ mainly consists of agreements then the ratio R would be large and thus the record pair would more likely to be designated as a link. On the other hand for a $\gamma \ \varepsilon \ \Gamma$ that primarily consists of disagreements