A Probabilistic Deduplication, Record Linkage and Geocoding System

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

Peter Christen\textsuperscript{1} and Tim Churches\textsuperscript{2}

\textsuperscript{1}Department of Computer Science, Australian National University, Canberra ACT 0200; peter.christen@anu.edu.au
\textsuperscript{2}Centre for Epidemiology and Research, NSW Department of Health, North Sydney NSW 2059; tchur@doh.health.nsw.gov.au

Abstract
Finding duplicate records in one data set, or linking records from two or more data sets are increasingly important tasks in the pre-processing phase of many data mining projects, as often information from multiple sources needs to be integrated, combined or linked in order to allow more detailed data analysis or mining. The aim of such linkages is to match all records relating to the same entities, such as patients. As common unique entity identifiers (or keys) are often not available, the linkage process needs to be based on the common attributes available in all the data sets to be linked, and becomes increasingly difficult with today’s large data collections.

Record linkage and deduplication 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. In the health sector, linked data might contain information that is needed to improve health policies, and that traditionally has been 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.

A related issue of increasing interest is the geocoding of records in order to allow spatial data analysis and mining, for example of disease outbreaks, or correlations with environmental factors. Geocoding is the process of matching a data source with geocoded reference data (which is made of cleaned and standardised records containing address information plus their geographical location). It is estimated that between 80\% and 90\% of governmental data collections contain address information.

In this paper we present an overview of the \textit{Febrl} (Freely extensible biomedical record linkage) project, which aims at developing improved algorithms and techniques for large scale data cleaning and standardisation, record linkage, deduplication and geocoding of massive data sets. We discuss new probabilistic techniques for data cleaning and standardisation, approximate geocode matching, parallelisation of blocking and linkage algorithms, as well as a probabilistic data set generator.

Freely available prototype software is available from the project Web site under an open source software license, allowing the modification, adjustment and improvement as needed by a user. \textit{Febrl} is an ideal platform for the rapid development, implementation, and testing of new and improved record linkage algorithms and techniques.