

Quality Assurance of Government Databases^{*}

Mohamed G. Elfeky

Ahmed K. Elmagarmid

Thanaa M. Ghanem

Department of Computer Sciences, Purdue University, West Lafayette, IN 47907.

{mgelfeky, ake, ghanemtm}@cs.purdue.edu

Abstract

Data cleaning is a vital process that ensures the quality of data stored in real-world databases. The process of identifying the record pairs that represent the same entity (duplicate records), commonly known as *record linkage*, is one of the essential elements of data cleaning. Digital government serves as an emerging area for database research, such as database management, data integration, data cleaning, etc. In this demo, we present a record linkage tool as part of a digital government web service. This demo serves as a framework for incorporating data quality tools with digital government web services.

1. Introduction

A typical and emerging area that involves access to both databases and applications is *Digital Government* [3]. The aim of digital government is to provide computer-based systems that allow dynamic management and access of a large number of governmental databases and services. The government data is so critical that it should be designed, analyzed and managed with *Data Quality* as a guiding principle and not as an afterthought [5].

Defined as “fitness for purpose”, data quality is multidimensional in the sense that it has several dimensions. The most common dimension is soundness: whether the data available contains the true values. Data soundness is often referred to as correctness, precision, accuracy or validity. Another common dimension of data quality is completeness: whether all the data are available. In the terminology of databases, data completeness refers to both the record completeness (no record is missing) and attribute completeness (no attribute value in a record is missing). Soundness and completeness are two orthogonal dimensions in the sense that they are concerned with completely separate aspects of data quality. Other important dimensions often discussed are consistency, currency and security.

There are two main concerns about data quality: data cleanup and process cleanup. Data cleanup is the current research concern. Quality of data can be improved significantly by identifying obvious errors in the data and by identifying records that correspond to the same entity. The former is referred to as “data editing”, and the latter is referred to as “record linkage”. Process cleanup is more important as it concerns

^{*} This research is partially supported by NSF under grant 9983249-EIA.

keeping the data clean. It focuses on measuring the quality dimensions and on trying to enforce these dimensions into data.

In this demo, we present a record linkage tool as part of a digital government web service that is provided for disadvantaged citizens. The rest of this demo description is organized as follows. In Section 2, the record linkage problem is introduced along with a brief description of the record linkage methodology. In section 3, we present the digital government web service, describing the system design and showing how to use the record linkage tool.

2. Record Linkage

Record linkage is the process of comparing the records from two or more data sources in an effort to determine which pairs of records represent the same real-world entity [2]. Record linkage may also be defined as the process of discovering the duplicate records in one file. What makes record linkage a problem in its own right, (i.e., different from the duplicate elimination problem), is the fact that real-world data is “dirty”. In other words, if data were accurate, record linkage would be similar to duplicate elimination, since the duplicate records would have the same values in all fields. Yet, in real-world data, duplicate records may have different values in one or more fields. For example, more than one record may correspond to the same person in a citizen database because of a misspelled character in the name field.

The record linkage process comprises two main steps. The first step is to generate the comparison vectors by component-wise comparison of each record pair. The second step is to apply the decision model to the comparison vectors to determine the matching status of each record pair. Figure 1 shows how the record linkage process operates. First, a searching method is exploited to reduce the size of the comparison space. It is very expensive to consider all possible record pairs for comparison. For a data file of n records, the number of record pairs that can be generated is equal to $n(n-1)/2$, i.e., $O(n^2)$. In order to reduce the large space of record pairs, searching methods are needed to select a smaller set of record pairs. The selected set of record pairs is called *reduced comparison space*. Since the main objective of the record linkage process is to detect the matched record pairs (duplicate records), searching methods try to select the record pairs that are candidates to be matched. They should be intelligent enough to exclude any record pair whose two records completely disagree, i.e., to exclude any record pair that cannot be a potentially matched pair. The selected record pairs are provided to the comparison functions to perform component-wise comparison of each record pair, and hence generate the comparison vectors. Then, the decision model is applied to predict the matching status of each comparison vector. Last, an evaluation step, to estimate the performance of the decision model, is performed. Table 1 gives a complete list of the various models and tools implemented in each component. We refer the interested reader to [2] for a detailed description about each of these models and tools.

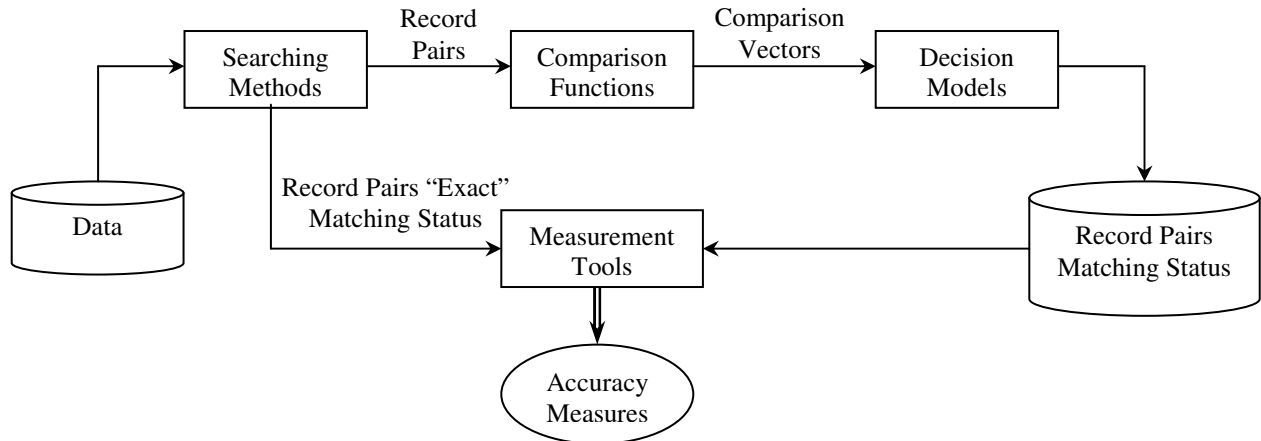


Figure 1. Record Linkage Process

Searching Methods	Comparison Functions	Decision Models	Measurement Tools	Supporting Tools
<ul style="list-style-type: none"> - Blocking - Sorting - Hashing - Sorted Neighborhood 	<ul style="list-style-type: none"> - Hamming Distance - Edit Distance - Jaro's Algorithm - N-grams - Soundex Code 	<ul style="list-style-type: none"> - Probabilistic Model - EM-Based - Cost-Based - Error-Based - Induction Model - Clustering Model - Hybrid Model 	<ul style="list-style-type: none"> - Reduction Ratio - Pairs Completeness - Accuracy - Completeness 	<ul style="list-style-type: none"> - MLC++ - ID3 decision trees - IBL instance-based learning - DBGen

Table 1. TAILOR Tools List

3. System Description

3.1 Web Services

Web Services is a very suitable approach that meets the needs of the governmental services. It provides the standard framework and protocols to implement, publish and communicate these services. Many new platforms and technologies appear to support web services; this will allow implementing different governmental services on different platforms and make efficient interaction between them. In other words, each government organization will be responsible for implementing its services according to a standardized framework, while the integration and interaction between these several organizations will be left to a web services *orchestration* manager.

3.2 Application Domain

Currently, disadvantaged citizens must collect their benefits by visiting several offices within and outside the towns in which they live [1]. To tackle this problem, we present a prototype for the *Independent Living* web service that can be provided by the government to disadvantaged citizens to maximize their integration in community leadership, independence and productivity. Mainly, this web service helps

them to get the benefits provided by the government easily without going to several offices. The IL (Independent Living) web service provides a *Programs* service that enables disadvantaged citizens to browse the programs provided in government-supported centers, and to register in them online. Such programs teach them how to get used to their disabilities. Moreover, the IL web service provides a *Housing* service that enables disadvantaged citizens to browse the houses provided by the government that meet their needs, and to make an online reservation for those houses.

Serving citizens, the IL web service uses a citizen database that is updated regularly through new citizens who register themselves in the web service, and through a database administrator who should regularly check the quality of this database. A record linkage tool is provided as part of the IL web service in order to help the administrator achieve this task.

3.2 Implementation

The system is built using *HP Netaction software suite* [4], which is a J2EE (Java 2 Enterprise Edition) platform for building web services. All the information related to the IL web service is stored in an *Oracle* database. Application programs are implemented in Java using JDBC (Java Database Connectivity) to connect to the Oracle database. *HP E-speak* middleware is used to host the application programs. An application server (*HP Total-e-Server*) connects to the e-speak core to invoke different methods of the application. The user interacts with the system through an interface implemented in JSP (Java Server Page) files that are hosted by an IIS (Internet Information Server) Web Server.

The database administrator can use the same interface to run the record linkage tool. The interface allows him to select a searching method, a comparison function, and a decision model, as well as to tune all the required parameters. The values of the parameters determine the functionality of the various components shown in Figure 1.

References

- [1] A. Bouguettaya, A.K. Elmagarmid, B. Medjahed, and M Ouzzani. A Web-based Architecture for Government Databases and Services. In *Proc. of National Conf. on Digital Government Research DG.O 2001*, Los Angeles, CA, May 2001.
- [2] M.G. Elfeky, V.S. Verykios, and A.K. Elmagarmid. TAILOR: A Record Linkage Toolbox. In *Proc. of Int. Conf. on Data Engineering ICDE'2002*, San Jose, CA, February 2002.
- [3] A.K. Elmagarmid and W.J. McIver. The Ongoing March Toward Digital Government. *IEEE Computer*, 34(2), February 2001.
- [4] HP Netaction Software Suite. <http://www.hp.com/products1/softwareproducts/software/netaction/>.
- [5] A. Umar, G. Karabatia, L. Ness, B. Horowitz, and A.K. Elmagarmid. Enterprise Data Quality: A Pragmatic Approach. *Information Systems Frontiers*, 1(3), 2000.