# Methods for Linking and Mining Massive Heterogeneous Databases

**José C. Pinheiro and Don X. Sun**
Statistics Research
Bell Laboratories, Lucent Technologies
600 Mountain Avenue, Murray Hill, NJ 07974
Email: jcp,dxsun@bell-labs.com

## Abstract

Many real-world KDD expeditions involve investigation of relationships between variables in different, heterogeneous databases. We present a dynamic programming technique for linking records in multiple heterogeneous databases using loosely defined fields that allow free-style verbatim entries. We develop an interestingness measure based on non-parametric randomization tests, which can be used for mining potentially useful relationships among variables. This measure uses distributional characteristics of historical events, hence accommodating variable-length records in a natural way. As an illustration, we include a successful application of the proposed methodology to a real-world data mining problem at Lucent Technologies.

## 1 Introduction

Many large scale data analysis problems involve the investigation of relationships between variables in heterogeneous databases with different temporal structures. For example, one may be interested in investigating relationships between customer satisfaction with products and services provided by a company and the company's in-house maintenance and sales records. Satisfaction surveys are generally conducted on a periodic basis and only involve a relatively small sample from the universe of customers. Maintenance and sales records, on the other hand, are collected continuously, providing massive amounts of information on all customers.

We concentrate on applications where two databases are to be combined and mined, but the methods described here extend easily to more than two databases. The scenario we consider consists of one database with massive amounts of data collected over time, with multiple records per individual and another, smaller database with a single record per individual. In Section 2, we describe a real-life example with this type of data structure referring to customer satisfaction data and maintenance records collected over several years at Lucent Technologies. This example is used throughout the paper to illustrate the different methods we present.

To link records with incomplete or missing common identifiers, we present in Section 3 a method for linking records in multiple heterogeneous databases, using loosely defined fields that allow free-style verbatim entries. A dynamic programming technique is applied to compute matching probabilities and the decision thresholds are estimated from some valid records known as training data. Section 4 briefly describes the strategies for collapsing and combining databases once record linkage has been established.

The combined database is used for mining interesting relationships among variables. Usually, a large number of variables is present in the data and it is desirable to employ automatic mining techniques to select a reasonable number of potentially interesting variables for further, more detailed investigation. We present, in Section 5, an interestingness measure, based on a non-parametric randomization test, which can be used for automatic data mining. We also describe graphical methods, based on Trellis displays (Becker, Cleveland, & Shyu 1996), for summarizing the results of the data mining search and for further exploring the relationships with largest interestingness values. These methods are model-free, robust to the presence of outliers, and scale-up to databases of arbitrary size. Our conclusions and suggestions for further research are included in Section 6.

## 2 An example

We introduce a real-life example that includes databases with different temporal structures which include a *customer satisfaction* database and a *maintenance service* database collected over the past several years at Lucent Technologies.

The customer satisfaction database contains records from a quarterly sample survey of Lucent Technologies' customers. The survey includes over twenty questions measuring customer satisfaction with various aspects of equipment and maintenance service. All questions use a 1–4 ordinal scale, with 1 meaning *very dissatisfied* and 4 *very satisfied*.

The maintenance database contains records pertaining to any maintenance service provided by Lucent over the past several years. Records are entered in this database whenever a new maintenance service is initiated or has its status modified, amounting to several gigabytes of data per month. Dozens of variables measuring different aspects of the maintenance service cycle are included in this database. Some examples are: *service duration*, *severity of the problem*, and *type of equipment involved.*

The objectives of the data mining investigation are, first, to verify if any relationships exist and, if so, to identify which customer satisfaction variables are more sensitive to maintenance service variables and which maintenance service variables most affect customer satisfaction. These can be used to determine potential areas of intervention for improving services to meet or exceed customer expectations.

# 3 Record linkage

In this section, we present a general method for matching verbatim text fields which is used to link records across different databases. First, we propose a text similarity measure between two sequences of characters based on a dynamic programming algorithm. The similarity measure ranges from 0 (indicating that the fields are completely dissimilar) to 1 (exact match). Based on the similarity measures for each corresponding pair of fields, we build a classification model using logistic regression to predict whether the two records are matched or not.

## 3.1 Text similarity measure

To find the best match of two text strings, we propose a text similarity measure. For a given text string, we use a vector to represent all the characters of the string with the consecutive space characters collapsed into one. For two given sequences $a_i, i = 1, \cdots, n$ and $b_i, i = 1, \cdots, m$, our objective is to find a map

$$M(\cdot) : \{1, \cdots, n\} \rightarrow \{1, \cdots, m, \emptyset\}$$

such that

$$\frac{\sum_{i=1}^{n} s(a_i, b_{M(i)})}{(n+m)/2}$$

is maximized. The map function satisfies the condition that $M(i) > M(j)$ for any pair of $(i, j)$ with $i > j$, $M(i) \neq \emptyset$, and $M(j) \neq \emptyset$. The character similarity function $s(\cdot, \cdot)$ is defined as

$$s(a_i, b_j) = \begin{cases} 1, & \text{if } a_i = b_j, j \neq \emptyset; \\ 0, & \text{otherwise.} \end{cases}$$

We define

$$S(\mathbf{a}, \mathbf{b}) = \max_{M(\cdot)} \frac{\sum_{i=1}^{n} s(a_i, b_{M(i)})}{(n+m)/2} \qquad (1)$$

as our text similarity measure.

The optimization problem in (1) can be solved using the well known dynamic time warping method. More details on the algorithm can be found in (Pinheiro & Sun 1998).

## 3.2 Prediction

Once the similarity measures are calculated for the corresponding fields of two records, they can be used as the basis for predicting whether the match is true or false. Let $x_1, \cdots, x_k$ be the variables representing the text similarity measures for all the verbatim fields that appear in both databases, and $y$ be the binary variable indicating if the match is actually true or false. We use a simple logistic regression model for this purpose:

$$\Pr(y = 1 | x_1, \cdots, x_k) = \frac{\exp(\beta_0 + \sum_{j=1}^{k} \beta_j x_j)}{1 + \exp(\beta_0 + \sum_{j=1}^{k} \beta_j x_j)} \qquad (2)$$

where $\beta_j, j = 0, \cdots, k$ are model parameters to be estimated from a given training data set.

Using this prediction model, we can link the records without common unique identifier in two databases A and B as follows. For each record in database B, we compute $x_{i1}, \cdots, x_{ik}$, the similarity measures between its $k$ text fields and the corresponding fields of the $i$-th record in database A. Then, we find the record in database A with the largest matching probability using (2): $i^* = \arg\max_{i \in A} \Pr(y = 1 | x_{i1}, \cdots, x_{ik})$. If

$$\Pr(y = 1 | x_{i^*1}, \cdots, x_{i^*k}) \geq \Pr(y = 0 | x_{i^*1}, \cdots, x_{i^*k}),$$

the record in database B is linked to the $i^*$-th record in database A. Otherwise, no link is established for this record between databases A and B.

In our customer satisfaction example, there are 500,000 records in the maintenance record database and 12,000 records in the customer survey database. Among these 12,000 records, only about 40% of the records do not have the unique identifier field. Four fields are chosen as the basis for matching records in this task: Customer Business Name: ($x_1$); Street Address: ($x_2$); City Name: ($x_3$); State Name: ($x_4$). To evaluate the proposed method of record linkage, we randomly split all the records with the common identifier field into two parts, one for training and the other for testing. We fit four different logistic regression models using various number of variables, and the result is shown in Table 1. In this example, using just one text field (*business name*) gives a satisfactory matching accuracy of 98%. An additional field of *street address* boosts the accuracy to over 99%.

| Model | Accuracy (Train) | Accuracy (Test) |
|---|---|---|
| $x_1$ | 98.5% | 98.7% |
| $x_1$+$x_2$ | 99.0% | 99.5% |
| $x_1$+$x_2$+$x_3$ | 99.3% | 99.8% |
| $x_1$+$x_2$+$x_3$+$x_4$ | 99.3% | 99.7% |

Table 1: Prediction results of record linkage based on text similarity measure of various text fields.

# 4 Combining databases

Because of the different temporal structures of the databases, individual records in the smaller database are generally linked to multiple records in the larger database. These multiple records need to be collapsed to generate a consolidated database with a single record per individual, which is used for mining potentially interesting relationships. Averages and medians are used to collapse numeric variables, while percentages and counts are used to replace categorical variables in the collapsed database. For example, in the customer satisfaction study introduced in Section 2, customer's maintenance service durations are summarized by the median service duration and the reporting status (whether or not a customer reported the problem) of the multiple maintenance records are represented by the percentage of customer reported problems.

## 5 Mining interesting relationships

This section describes a methodology for screening potentially interesting relationships, based on a model-free interestingness measure and Trellis graphical displays.

### 5.1 An interestingness measure

We denote a generic variable in the collapsed massive database by $X$ and assume that it takes numeric values. A generic variable in the smaller database with unique records per individual is denoted by $Y$ and, without loss of generality, we assume that it takes values on a discrete set. We denote by $n_y$ the number of possible values of $Y$.

A way of characterizing how *interesting* is the relationship between $X$ and $Y$ is by measuring how much the conditional distribution of $X$ given $Y$ differs from the marginal distribution of $X$, that is, how much knowing that $Y = y$ affects the chances of $X$ taking a value $x$.

A non-parametric description of the distribution of $X|Y = y$ is provided by the *quantiles* of that distribution (Conover 1980, p. 29). Similarly, the empirical conditional distribution of $X|Y = y$ may be described by the sample quantiles of the values $X$ that were observed with $Y = y$. Generally, only a small number of quantiles, $n_q$, are required to give a good representation of the distribution. If $X$ and $Y$ are independent, the quantiles of $X|Y = y$ are independent of $y$. The amount by which the quantiles of $X|Y = y$ vary with $y$ relates to the interestingness of the underlying relationship. Because the *true* quantiles are not known, the empirical quantiles are used to evaluate the differences in the conditional distributions.

Let $E_p$ denote the set of equivalent empirical quantiles associated with the different values of $y$, corresponding to a probability $p$ (e.g. all 25% quantiles). Under the assumption that $X$ and $Y$ are independent, all elements in $E_p$ estimate the same theoretical quantile and any ordering of them is equally likely to be observed. Replacing the actual values of the empirical quantiles by their respective ranks within $E_p$, it follows that, under the assumption of independence, all $n_y!$ rank permutations are equally likely. By convention, ties are assigned the average rank of the elements involved. Intraclass ranks for the different $E_p$ quantile classes can be independently permuted, resulting in $N(X,Y) = (n_y!)^{n_q}$ equally likely permutations. These can be used to derive a reference distribution for measuring the distance between the conditional distributions and, hence, the interestingness.

Let $R_{ij}$ denote the rank of the $j$th empirical quantile corresponding the $i$th value of $y$ within its $E_p$ class. The following statistics can be used to measure the difference between the conditional distributions.

$$\begin{aligned} K(X,Y) &= \sum_{i=1}^{n_y} (R_{i.} - R_{..})^2 \,/ \sum_{i=1}^{n_y} \sum_{j=1}^{n_q} (R_{ij} - R_{i.})^2 \\ R_{i.} &= \sum_{j=1}^{n_q} R_{ij}/n_q, \quad R_{..} = \sum_{i=1}^{n_y} R_{i.}/n_y \qquad (3) \end{aligned}$$

$K$ is similar to a Kruskal–Wallis non-parametric test statistics (Conover 1980, p. 229) for testing equality of means. Intuitively, if the conditional distributions present some sort of stochastic ordering, there will be an association between the empirical ranks and $y$, leading to larger deviations between average ranks (the numerator of $K$) and smaller within-class deviations (the denominator of $K$), resulting in larger values of $K$. A reference distribution for $K$ can be constructed by considering permutations $\pi$ of the ranks within each $E_p$ class and applying (3) to the permuted ranks to obtain a new value $K_\pi$ (Good 1995). This reference distribution can be used to calculate a randomization test p-value for the observed $K$ (Good 1995), which constitute our interestingness measure for the pair $(X,Y)$

$$\alpha(X,Y) = \sharp \left\{ \pi : K_\pi \geq K(X,Y) \right\} / N(X,Y) \qquad (4)$$

That is, $\alpha(X,Y)$ gives the percentage of $K_\pi$ that are greater than or equal to $K(X,Y)$. $\alpha$ can be interpreted as a p-value for the null hypothesis that $X$ and $Y$ are independent and the smaller the value of $\alpha$, the more *interesting* the relationship between $X$ and $Y$. Because $\alpha$ is derived from a randomization test based on intraclass ranks of quantiles, it is model-free and robust to the presence of outliers. Note, in particular, that $\alpha$ is invariant to $1-1$ transformations of X. Also, because the number of quantiles $n_q$ can be kept fixed, it scales-up to arbitrarily large databases. When $N(X,Y)$ is too large for complete enumeration of the reference set, a large sample of random permutations is used to estimate $\alpha$ (Good 1995).

### 5.2 Exploring interestingness

We consider the customer satisfaction example of Section 2 to illustrate the use of the interestingness measure $\alpha$ described in Section 5.1. There are a total of 21 $Y$ variables in the customer satisfaction database, all measured on an ordinal 1–4 scale ($n_y = 4$), and 13 numeric $X$ variables in the consolidated maintenance service database, corresponding to 273 $(X,Y)$ pairs. The 10%, 25%, 50%, 75%, and 90% quantiles are used to represent the empirical conditional distributions of $X|Y$ ($n_q = 5$).

As a first application of $\alpha$, we consider the problem of determining a time window for collapsing the maintenance records. As described in Section 4, because the two databases have different temporal structures, records in the larger database need to be collapsed over a time window. This window must include the quarter in which the associated record in the customer data was collected, but its width may vary. Short windows may lead to a loss of relevant records, but long windows may include data no longer associated with the customer's responses. That may also vary with both $X$ and $Y$. A total of 8 widths, ranging from 0 to 36 months are considered for the customer satisfaction example.

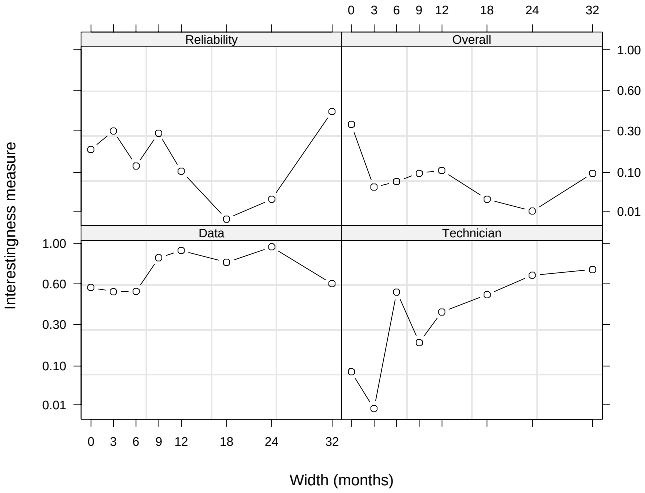

Figure 1: Interestingness of four customer satisfaction variables with respect to median duration of maintenance service, versus time window width (in months).

Figure 1 gives a Trellis display of $\alpha$ versus window width for the $Y$ variables *Data* (appropriateness of data), *Technician* (technician's knowledge), *Reliability*, and *Overall* (overall satisfaction), with respect to the median maintenance service duration. Each panel of the trellis corresponds to a different $Y$ and the same scale is used in all panels, to facilitate their comparison. A square-root scale is used for $\alpha$ to enhance visualization. *Data* does not seem to be related to service duration, as all its $\alpha$ values are above 0.30. The highest interestingness value for *Technician* occurs for a 3-month window, suggesting that this is a "short memory" variable. The optimal widths for *Reliability* and *Overall* are respectively 18 and 24 months, suggesting that these are "long memory" variables. All of these last three variables show potentially interesting relationships with service duration, which should be further investigated. Similar analyses are done for the other $(X, Y)$ pairs.

The analysis objectives for the customer satisfaction project are to determine which maintenance service variables have greater impact on customer satisfaction and which customer satisfaction variables are most sensitive to maintenance service performance. The interestingness measure $\alpha$ can be used to address both of these issues. Figure 2 gives a Trellis display of the minimum $\alpha$ over window widths for a subset of maintenance and customer satisfaction variables. The square-root scale is used again.

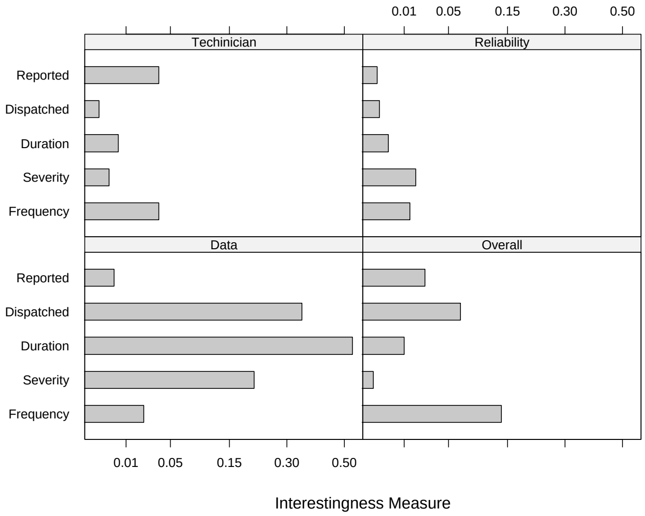

Figure 2: Interestingness measures for a subset of maintenance service and customer satisfaction variables, with different panels for the customer satisfaction variables.

It is clear that *Data* and *Overall* are less sensitive to maintenance service variables than *Technician* and *Reliability*. *Frequency* of problems seems to be the less influential maintenance variable, but this becomes clearer in the Trellis display of Figure 3, where the panels are now determined by the maintenance variables and the rows within the panels by the customer satisfaction variables.

Figure 3 also reveals that *Reported* (percentage of times a problem was reported by the customer) has uniformly low $\alpha$ values, suggesting it is an influential variable on customer satisfaction.

### 5.3 Understanding Interestingness

The sample distribution of a variable is compactly represented by its *boxplot* (Velleman & Hoaglin 1981). Boxplots have good scalability properties, because they are based on a few quantiles of the data.

Comparison of the conditional distributions of $X|Y = y$ is done by plotting, side by side, the boxplots corresponding to each $y$. Trellis displays provide a powerful graphical environment for combining several of these plots, facilitating their comparison and understanding. Figure 4 gives an example of such a Trellis display for the customer satisfaction data. The $Y$ variables are *Data*, *Overall*, *Reliability*, and *Technician*,

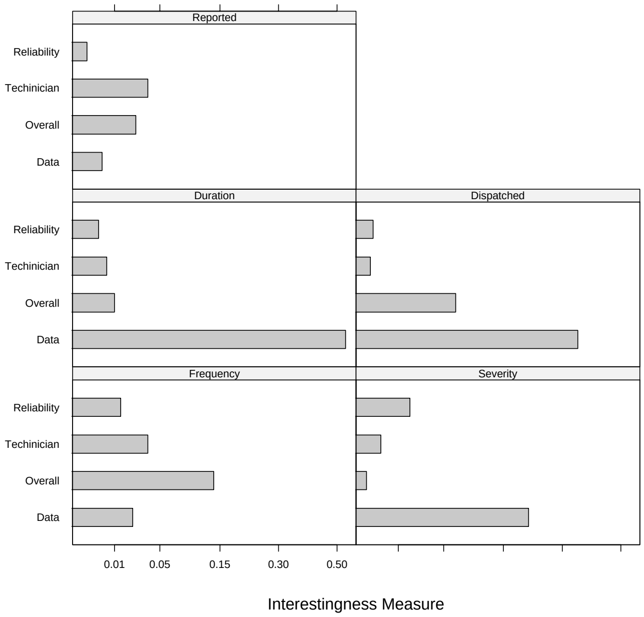

Figure 3: Interestingness measures for the same subset of variables as in Figure 2, with different panels for the maintenance variables.

each corresponding to a different panel, and the $X$ variable is *Reported*.

It is clear from Figure 4 that customers with *poor* satisfaction levels have different *Reported* values than the rest of the customers. This is more evident for the variables *Overall* and *Reliability*, for which the dissatisfied customers present higher values of reported problems. This information is useful for identifying and prioritizing customer satisfaction problem areas.

Similar Trellis displays are used for all potentially interesting relationships flagged by the interestingness measure. Human intervention is required at this stage of the analysis to sort out the relevant relationships and to decide on their usefulness. The automatic screening of potentially interesting relationships prior to this step greatly reduces the need of human intervention, allowing the most valuable resources to be selectively allocated.

## 6 Discussion

We develop a methodology for linking, combining, and mining massive heterogeneous databases. We propose a method for linking records in multiple heterogeneous databases using loosely defined fields that allow free-style verbatim entries. A dynamic programming technique is developed to compute matching probabilities, with the decision thresholds being estimated from training data. To screen potentially interesting relationships between variables in the massive databases, we present an interestingness measure based on a non-parametric randomization test, which can be used for automatic data mining. We describe graphical methods, based on Trellis displays, for summarizing the results of the data mining search and for further exploring the relationships with largest interestingness values. These methods are model-free, robust to the presence of outliers, and scale-up to databases of arbitrary size.

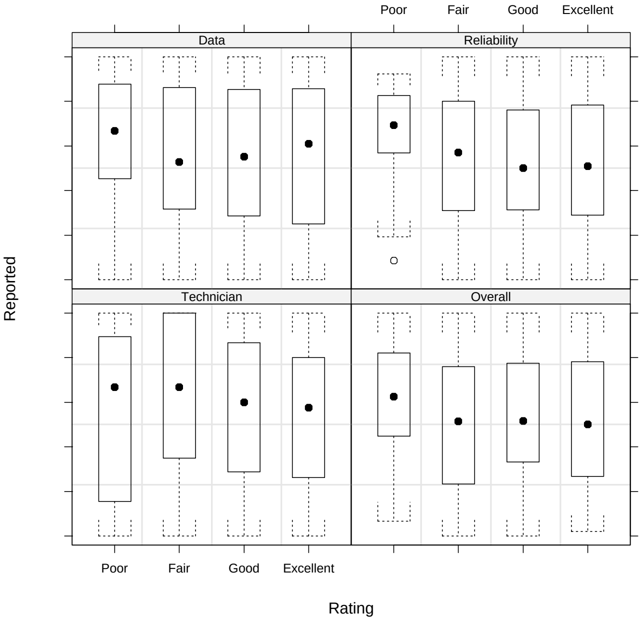

Figure 4: Boxplots of the sample distributions of *Reported* conditional on satisfaction ratings for a subset of the customer satisfaction variables.

## 7 Acknowledgements

We would like to thank Lorraine Denby and James M. Landwehr from the Statistics Research group at Bell Labs for their helpful comments and suggestions.

## 8 References

Becker, R. A.; Cleveland, W. S.; and Shyu, M.-J. 1996. The visual design and control of trellis graphics displays. *J. of Computational and Graphical Statistics* 5(2):123–156.

Conover, W. J. 1980. *Practical Nonparametric Statistics.* New York: Wiley, 2nd edition.

Good, P. 1995. *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses.* New York: Springer–Verlag.

Pinheiro, J. C., and Sun, D. X. 1998. Methods for linking and mining massive heterogeneous databases. Technical memorandum, Bell Laboratories, Lucent Technologies.

Velleman, P., and Hoaglin, D. 1981. *Applications, basics, and computing of exploratory data analysis.* New York: Duxbury Press.