United States™ Census Bureau

NORC at the UNIVERSITY of CHICAGO

Abstracts for the

Person Validation and Entity Resolution Conference

May 23rd, 2011, Washington D.C.

United States™ Census Bureau

NORC at the UNIVERSITY of CHICAGO

Contents

Session I: International Papers

GLink, A Probabilistic Record Linkage System 4

Antoine Chevrette and Michael Wenzowski, Statistics Canada

Indigenous Life Expectancy using Multiple Australian Data Sources 5

Richard Madden, University of Sydney; Leonie Tickle, Macquarie University; Lisa Jackson-Pulver, University of New South Wales; Ian Ring, University of Wollongong

A Hit-Miss Model for Duplicate Detection in the WHO Drug Safety Database 5

Niklas Norén, Roland Orre, Stockholm University, and Andrew Bate, Pfizer Inc.

Session II: Fuzzy Matches and Blocking

Application of probabilistic linkage methods to join infectious disease surveillance records to death registrations. 6

Theresa L. Lamagni; Nicola Potz, David Powell, Nicholas Hinton; Andrew Grant; Elizabeth Sheridan; Richard Pebody, Health Protection Agency, U.K

Error-Tolerant Record Matching 7

Surajit Chaudhuri, Microsoft, Venkatesh Ganti, Google, and Rajeev Motwani, Stanford University

Identifying Value Mappings for Data Integration: An Unsupervised Approach 7

Jaewoo Kang, Korea University; Dongwon Lee, The Pennsylvania State University; and Prasenjit Mitra, The Pennsylvania State University

Learning Blocking Schemes for Record Linkage 7

Matt Michelson, Fetch Technologies, and Craig Knoblock, University of Southern California

Session III: Dynamic Processing

Collective Entity Resolution 8

Lise Getoor, University of Maryland, College Park, and Indrajit Bhattacharya, IIS Bangalore

A Case Study in Record Linkage 8

Marianne Winglee, Westat; Richard Valliant, University of Michigan; and Fritz Scheuren, NORC at the University of Chicago

Conflicting data: The Role of Source Dependence 9

Xin Luna Dong, Laure Berti-Equille, and Divesh Srivastava, AT&T

SESSION I: INTERNATIONAL PAPERS

GLINK, A PROBABILISTIC RECORD LINKAGE SYSTEM

Antoine Chevrette and Michael Wenzowski, Statistics Canada

Abstract

It is commonly the case that data originating from disparate sources, and even from the same source over time, exists in incompatible formats. The consequence is that the analysis of such data must often be preceded by some form of record linkage operation in order to assemble the data into a well-structured form. When reliable and invariant identifiers are available, such linkages become a relatively easy task to perform. In other cases, the identifier values may vary, requiring the reformatting of the data and possibly the execution of a fuzzy match. In both of these scenarios, "off the shelf" software is readily available that is well-suited to the task at hand. However, in cases in which a greater degree of variance must be accommodated, especially in cases in which multiple fields must be examined in order to identify a linkage, the task falls to software capable of performing a sophisticated and complex probabilistic record linkage. Unfortunately, software capable of this degree of sophistication is very highly specialized and only sparsely available.

We present the results of a recent Statistics Canada initiative to re-engineer our generalized record linkage system in order to enhance its applicability across a wide range of processing problem and subject matter domains. The software faithfully implements the probabilistic record linkage methodology first described by Fellegi and Sunter, and includes many extensions and enhancements to increase the utility of the application. We will demonstrate how we have improved the software by offering more intuitive controls over managing the complexity of internal processing; by extending and enhancing the software's capabilities; and by simplifying the installation, setup and processing models.

Indigenous Life Expectancy using Multiple Australian Data Sources

Richard Madden, University of Sydney; Leonie Tickle, Macquarie University; Lisa Jackson-Pulver, University of New South Wales; Ian Ring, University of Wollongong

Abstract

In 2009, the Australian Bureau of Statistics (ABS) released new estimates of Indigenous life expectancy for Australia. The estimates, which were substantially higher than previously published estimates, were based on linkage between Indigenous deaths registered in the period from August 7, 2006 to June 30, 2007 and 2006 census records. State estimates were also produced for some states, showing substantial variations between states; the life expectancies are inversely related to the calculated completeness of Indigenous identification of death registrations. Analysis based on more comprehensive linkage of death records in New South Wales (NSW) over 5 years suggests that the ABS methods have understated Indigenous deaths and so overstated life expectancy.

The paper will report the NSW results, based on several linkage algorithms, including a comparison of the algorithms. Resulting changes in life expectancy estimates will be reported. Suggested improvements to ABS methods will be discussed.

A Hit-Miss Model for Duplicate Detection in the WHO Drug Safety Database

Niklas Norén, Roland Orre, Stockholm University, and Andrew Bate, Pfizer Inc.

Abstract

The WHO Collaborating Centre for International Drug Monitoring in Uppsala, Sweden, maintains and analyses the world's largest database of reports on suspected adverse drug reaction incidents that occur after drugs are introduced on the market. As in other post-marketing drug safety data sets, the presence of duplicate records is an important data quality problem and the detection of duplicates in the WHO drug safety database remains a formidable challenge, especially since the reports are anonymised before submitted to the database. However, to our knowledge no work has been published on methods for duplicate detection in post-marketing drug safety data. In this paper, we propose a method for probabilistic duplicate detection based on the hit-miss model for statistical record linkage described by Copas & Hilton. We present two new generalisations of the standard hit-miss model: a hit-miss mixture model for errors in numerical record fields and a new method to handle correlated record fields. We demonstrate the effectiveness of the hit-miss model for duplicate detection in the WHO drug safety database both at identifying the most likely duplicate for a given record (94.7% accuracy) and at discriminating duplicates from random matches (63% recall with 71% precision). The proposed method allows for more efficient data cleaning in post-marketing drug safety data sets, and perhaps other applications throughout the KDD community.

SESSION II: Fuzzy Matching and Blocking

Application of probabilistic linkage methods to join infectious disease surveillance records to death registrations

Theresa L. Lamagni; Nicola Potz, David Powell, Nicholas Hinton; Andrew Grant; Elizabeth Sheridan; Richard Pebody, Health Protection Agency, U.K.

Abstract

The sharing of health datasets between organizations offers unprecedented opportunities for addressing important public health questions through the integration of valuable information held solely by each organization. Joining large and disparate datasets does however raise concerns over the safeguarding of confidential records as well as presenting methodological challenges where unique and error-free personal identifiers are not universally available. A study was commissioned by the Department of Health to examine mortality in patients diagnosed with meticillin-resistant *Staphylococcus aureus* (MRSA) bacteraemia in England in 2004-05. Executing the study required linkage of laboratory data held by the Health Protection Agency (n=10,305) to statutory death registrations held by the Office for National Statistics (n=1,153,221). Given the low completion (<30%) of unique patient identifier (National Health Service number) but the availability of other personal identifiers within the MRSA dataset, a probabilistic record linkage method was developed. Data were dual-blocked on date of birth and soundex (code derived from surname) and linked using identifiers available (NHS number, forename initial, soundex, sex, date of birth, postal code). The linkage mechanism was evaluated through use of an independent source of information on patient outcome (NHS Central Register) and found to have high positive and negative predictive values (>90%) for correctly matching records. The linked data were therefore used to assess case fatality rates following MRSA infection and since applied to link other health datasets. The availability of methodological innovations facilitating complex data linkage within a climate of interagency cooperation and legislative support provide powerful and cost-effective tools for the advancement of scientific frontiers. Obtaining and retaining public confidence and support for such initiatives by demonstrating the potential public benefit will be a critical factor determining the future of this approach.

Error-Tolerant Record Matching

Surajit Chaudhuri, Microsoft; Venkatesh Ganti, Google; and Rajeev Motwani

Abstract

Record Matching is a key element of data cleaning technology. Error-Tolerant Record Matching reconciles multiple representations of the same entity in the presence of errors such as spelling mistakes and abbreviations. In this talk, we describe some of the key scenarios and the underlying technology for error-tolerant record matching that we have developed as part of our Data Cleaning project at Microsoft Research.

Identifying Value Mappings for Data Integration: An Unsupervised Approach

Jaewoo Kang, Korea University; Dongwon Lee, The Pennsylvania State University; and Prasenjit Mitra, The Pennsylvania State University

Abstract

The Web is a distributed network of information sources where the individual sources are autonomously created and maintained. Consequently, syntactic and semantic heterogeneity of data among sources abound. Most of the current data cleaning solutions assume that the data values referencing the same object bear some textual similarity. However, this assumption is often violated in practice. "Two-door front wheel drive" can be represented as "2DR-FWD " or "R2FD", or even as "CAR TYPE 3 " in different data sources. To address this problem, we propose a novel two-step automated technique that exploits statistical dependency structures among objects which is invariant to the tokens representing the objects. The algorithm achieved a high accuracy in our empirical study, suggesting that it can be a useful addition to the existing information integration techniques.

Learning Blocking Schemes for Record Linkage

Matt Michelson, Fetch Technologies, and Craig Knoblock, University of Southern California

Abstract

Record linkage is the process of matching records across data sets that refer to the same entity. One issue within record linkage is determining which record pairs to consider, since a detailed comparison between all of the records is impractical. Blocking addresses this issue by generating candidate matches as a preprocessing step for record linkage. For example, in a person matching problem, blocking might return all people with the same last name as candidate matches. Two main problems in blocking are the selection of attributes for generating the candidate matches and deciding which methods to use to compare the selected attributes. These attribute and method choices constitute a blocking scheme. Previous approaches to record linkage address the blocking issue in a largely ad-hoc fashion. This paper presents a machine learning approach to automatically learn effective blocking schemes. We validate our approach with experiments that show our learned blocking schemes outperform the ad-hoc blocking schemes of non-experts and perform comparably to those manually built by a domain expert.

SESSION III: DYNAMIC PROCESSING

COLLECTIVE ENTITY RESOLUTION

Lise Getoor, University of Maryland, College Park, and Indrajit Bhattacharya, IIS Bangalore

Abstract

In many domains, entity resolution results can be enhanced by combining information about the entity's attributes, together with co-occurrence information about the entities. We have developed two approaches for collective entity resolution, a relational clustering algorithm and a probabilistic approach based on the Latent-Dirichlet Allocation model. In both cases, the entity resolution decisions are not considered on an independent pairwise basis, but instead decisions are made collectively and we focus on how the use of relational links among the references can be exploited. We compare and contrast the two approaches, and demonstrate their utility on three real-world bibliographic datasets

A CASE STUDY IN RECORD LINKAGE

Marianne Winglee, Westat; Richard Valliant, University of Michigan; and Fritz Scheuren, NORC at the University of Chicago

Abstract

Record linkage is a process of pairing records from two files and trying to select the pairs that belong to the same entity. The basic framework uses a match weight to measure the likelihood of a correct match and a decision rule to assign record pairs as "true" or "false" match pairs. Weight thresholds for selecting a record pair as matched or unmatched depend on the desired control over linkage errors. Current methods to determine the selection thresholds and estimate linkage errors can provide divergent results, depending on the type of linkage error and the approach to linkage. This paper presents a case study that uses existing linkage methods to link record pairs but a new simulation approach (SimRate) to help determine selection thresholds and estimate linkage errors. SimRate uses the observed distribution of data in matched and unmatched pairs to generate a large simulated set of record pairs, assigns a match weight to each pair based on specified match rules, and uses the weight curves of the simulated pairs for error estimation.

Integrating Conflicting Data: The Role of Source Dependence

Xin Luna Dong, Laure Berti-Equille, and Divesh Srivastava, AT&T

Abstract

Many data management applications, such as setting up Web portals, managing enterprise data, managing community data, and sharing scientific data, require integrating data from multiple sources. Each of these sources provides a set of values and different sources can often provide conflicting values. To present quality data to users, it is critical that data integration systems can resolve conflicts and discover true values. Typically, we expect a true value to be provided by more sources than any particular false one, so we can take the value provided by the majority of the sources as the truth. Unfortunately, a false value can be spread through copying and that makes truth discovery extremely tricky. In this paper, we consider how to find true values from conflicting information when there are a large number of sources, among which some may copy from others.

We present a novel approach that considers dependence between data sources in truth discovery. Intuitively, if two data sources provide a large number of common values and many of these values are rarely provided by other sources (e.g., particular false values), it is very likely that one copies from the other. We apply Bayesian analysis to decide dependence between sources and design an algorithm that iteratively detects dependence and discovers truth from conflicting information. We also extend our model by considering accuracy of data sources and similarity between values. Our experiments on synthetic data as well as real-world data show that our algorithm can significantly improve accuracy of truth discovery and is scalable when there are a large number of data sources.