Data-Centric Systems and Applications

For further volumes:
http://www.springer.com/series/5258

Peter Christen

Data Matching

Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection

Springer

Peter Christen
Research School of Computer Science
The Australian National University
Canberra, ACT
Australia

ISBN 978-3-642-31163-5 ISBN 978-3-642-31164-2 (eBook)
DOI 10.1007/978-3-642-31164-2
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012942028

ACM Computing Classification (1998): H.2, H.3, I.2, I.5

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To Gail

Foreword

Early record linkage was often in the health area where individuals wanted to link patient medical records for certain epidemiological research. We can imagine the difficulty of comparing quasi-identifying information such as name, date of birth, and other information from a single record against a large stack of paper records. To facilitate the matching, someone might transfer the quasi-identifying information from a set of records to a large typed list on paper and then, much more rapidly, go through the large list. Locating matching pairs increases in difficulty because individual records might have typographical error ('Jones' versus 'Janes', 'March 17, 1922' versus 'March 27, 1922' because handwriting was difficult to read). Additional errors might occur during transcription to the typewritten list.

Howard Newcombe, a geneticist, introduced the idea of odds ratios into a formal mathematics of record linkage. The idea was that less frequent names such as 'Zbigniew' and 'Zabrinsky' (in English speaking countries) had more distinguishing power than more common names such as 'John' and 'Smith'. Among a pair of records that were truly matches, it was more typical to agree on several quasi-identifying fields such as first name, day of birth, month of birth, and year of birth than among a pair of records that had randomly been brought together from two files.

Newcombe's ideas were formulated in two seminal papers (Science, 1959; Communications of the Association of Computing Machinery, 1962). These papers contained a number of practical examples on combining the scores (odds ratios) from comparisons of individual fields in a pair to a total score (or total matching weight) associated with a pair. The combining of the logarithms of the scores via simple addition is under conditional independence (or näıve Bayes in machine learning). Pairs above a certain higher cutoff score were designated as links (or matches); pairs below a certain lower score were designated as a non-link (or non-match); and pairs between the upper and lower cutoff scores were known as potential links (potential matches) and held for clerical review. During the clerical review, the clerk might correct a name or date of birth by consulting an alternative source (list) or the original form (that might have had typographical error introduced during data capture to the computer files).

Obtaining the odds–ratios for suitably high quality matching would have been very difficult in most situations because training data were never available. Newcombe had the crucial insight that it was possible to compute the desired probabilities from large national files such as health or death indexes (or even censuses). He obtained the probabilities associated with the linked pairs by summing the valuespecific frequencies for individual first names, last names, etc., from the large file. He used the frequencies from the cross-product of the files along with an adjustment for those frequencies associated with linked pairs to get the appropriate frequencies for non-linked pairs. Newcombe's methods were robust with new pairs of files because the 'absolute' frequencies from the large national files worked well.

Fellegi and Sunter (Journal of the American Statistical Association, 1969) provided a formal mathematical model where they proved the optimality of Newcombe's rules under fixed upper bounds on the false link (match) rates and the false non-link (non-match) rates. The methods were later rediscovered by Cooper and Maron (Journal of the Association of Computing Machinery, 1977) without proofs of optimality. Fellegi and Sunter extended the model with ideas of unsupervised learning and extensions of value-specific frequency concepts with crude (but effective) ideas for typographical error rates.

For more than a decade, most of the methodological research has been in the computer science literature. Active areas are concerned with significantly improving linking speed with parallel computing and sophisticated retrieval algorithms, improving matching accuracy with better machine learning models or third-party auxiliary files, estimating error rates (often without training data), and adjusting statistical analyses in merged files to account for matching error.

Many applications are still in the epidemiological or health informatics literature with most individuals using government health agency shareware based on the Fellegi–Sunter model. Although individuals have introduced alternative classification methods based on Support Vector Machines, decision trees and other methods from machine learning, no method has consistently outperformed methods based on the Fellegi–Sunter model, particularly with large day-to-day applications with tens of millions of records.

Within this framework of historical ideas and needed future work, Peter Christen's monograph serves as an excellent compendium of the best existing work by computer scientists and others. Individuals can use the monograph as a basic reference to which they can gain insight into the most pertinent record linkage ideas. Interested researchers can use the methods and observations as building blocks in their own work. What I found very appealing was the high quality of the overall organization of the text, the clarity of the writing, and the extensive bibliography of pertinent papers. The numerous examples are quite helpful because they give real insight into a specific set of methods. The examples, in particular, prevent the researcher from going down some research directions that would often turn out to be dead ends.

Suitland, USA

William E. Winkler
U. S. Census Bureau

Preface

Objectives

Data matching is the task of identifying, matching, and merging records that correspond to the same entities from several databases. The entities under consideration most commonly refer to people, such as patients, customers, tax payers, or travellers, but they can also refer to publications or citations, consumer products, or businesses. A special situation arises when one is interested in finding records that refer to the same entity within a single database, a task commonly known as *duplicate detection*. Over the past decade, various application domains and research fields have developed their own solutions to the problem of data matching, and as a result this task is now known by many different names. Besides data matching, the names most prominently used are *record* or *data linkage*, *entity resolution*, *object identification*, or *field matching*.

A major challenge in data matching is the lack of common entity identifiers in the databases to be matched. As a result of this, the matching needs to be conducted using attributes that contain partially identifying information, such as names, addresses, or dates of birth. However, such identifying information is often of low quality. Personal details especially suffer from frequently occurring typographical variations and errors, such information can change over time, or it is only partially available in the databases to be matched.

There is an increasing number of application domains where data matching is being required, starting from its traditional use in the health sector and national censuses (two domains that have applied data matching for several decades), national security (where data matching has become of high interest since the early 2000s), to the deduplication of business mailing lists, and the use of data matching more recently in domains such as online digital libraries and e-Commerce.

In the past decade, significant advances have been achieved in many aspects of the data matching process, but especially on how to improve the accuracy of data matching, and how to scale data matching to very large databases that contain many millions of records. This work has been conducted by researchers in various

fields, including applied statistics, health informatics, data mining, machine learning, artificial intelligence, information systems, information retrieval, knowledge engineering, the database and data warehousing communities, and researchers working in the field of digital libraries. As a result, a variety of data matching and deduplication techniques is now available. Many of these techniques are aimed at specific types of data and applications. The majority of techniques has only been evaluated on a small number of (test) data sets, and so far no comprehensive large-scale surveys have been published that evaluate the various data matching and deduplication techniques that have been developed in different research fields.

The diverse and fragmented publication of work conducted in the area of data matching makes it difficult for researchers to stay at the forefront of developments and advances on this topic. This is especially the case for graduate and research students entering this area of research. There are no dedicated conferences or journals where research in data matching is being published. Rather, research in this area is disseminated in data mining, databases, knowledge engineering, and other fields as listed above. For practitioners, who aim to learn about the current state-of-theart data matching concepts and techniques, it is difficult to identify work that is of relevance to them.

While there is a large number of research publications on data matching available in journals as well as conference and workshop proceedings, thus far only a few books have been published on this topic. Newcombe [199] in 1988 covered data matching from a statistical perspective, and how it can be applied in domains such as health, statistics, administration, and businesses. Published at around the same time, the edited book by Baldwin et al. [16] concentrated on the use of data matching in the medical domain. More recently, Herzog et al. [143] discussed data matching as being one crucial technique required for improving data quality (with data editing being the second technique). A similar approach was taken by Batini and Scannapieco [19], who covered data matching in one chapter of their recent book on data quality. While Herzog et al. approach the topic from a statistical perspective, Batini and Scannapieco discuss it from a database point of view. Published in 2011, the book by Talburt [249] discusses data matching and information quality, and presents both commercial as well as open source matching systems. Similarly, Chan et al. [51] present declarative and semantic data matching approaches in several chapters in their recent book on data engineering.

None of these books however cover data matching in both the depth and breath this topic deserves. They either present only a few existing techniques in detail, provide a broad but brief overview of a range of techniques, or they discuss only certain aspects of the data matching process. The objectives of the present book are to cover the current state of data matching research by presenting both concepts and techniques as developed in various research fields, to describe all aspects of the data matching process, and to cover topics (such as privacy issues related to data matching) that have not been discussed in other books on data matching.

Organization

This book consists of 10 chapters. Chapter 1 provides an introduction to data matching (including how data matching fits into the broader topics of data integration and link analysis), a short history of data matching, as well as a series of example applications that highlight the importance and diversity of data matching. Chapter 2 then gives an overview of the data matching process and introduces the major steps of this process. A small example is used to illustrate the different aspects and challenges involved in each of these steps.

The core of the book is made of Chaps. 3–7. Each of these chapters is dedicated to one of the major steps of the data matching process. They each present detailed descriptions of both traditional and state-of-the-art techniques, including recently proposed research approaches. Advantages and disadvantages of the various techniques are discussed. Each chapter ends with a section on practical aspects that are of relevance when data matching is employed in real-world applications, and with a section on open problems that can be the basis for future research.

Chapter 3 discusses the importance of data pre-processing (data cleaning and standardising), which often has to be applied to the input databases prior to data matching in order to achieve matched data of high quality. The topic of Chap. 4 is the different indexing (also known as blocking) techniques that are aimed at reducing the quadratic complexity of the naive process of pair-wise comparing each record from one database with all records in the other database. The actual comparison of records and their attribute (or field) values is then covered in Chap. 5, with an emphasis put on the various approximate string comparison techniques that have been developed. How to accurately classify the compared record pairs into matches and non-matches is then discussed in Chap. 6. Both supervised and unsupervised classification techniques, and pair-wise and collective techniques are presented. Finally, Chap. 7 describes how to properly evaluate the quality and complexity of a data matching exercise. This chapter also covers the manual clerical review process that traditionally has been (and commonly still is) used within certain data matching systems, and the various publicly available test data collections and data generators that can be of value to both researchers and practitioners.

The final part of the book then covers additional topics, starting in Chap. 8 with a discussion of the privacy aspects of data matching, which can be of importance because personal information is commonly required for matching data. This chapter also provides an overview of recent work into privacy-preserving data matching (how databases can be matched without any private or confidential information being revealed). Chapter 9 presents a series of topics that can be of interest to both practitioners as well as the data matching research community. These topics include matching geo-spatial data, matching unstructured or complex types of data, matching data in real-time, matching dynamic databases, and conducting data matching on parallel and distributed computing platforms. This chapter also includes a list of open research topics. Finally, the book concludes in

Chap. 10 with a checklist of how data matching systems can be evaluated, and a brief overview of several freely available data matching systems.

Rather than providing definitions of relevant terms and concepts throughout the book, a glossary is provided at the end of the book (on page 243 onwards) that can help the reader to access the terms and concepts they are unfamiliar with.

Intended Audience

The aim of this book is to be accessible to researchers, graduate and research students, and to practitioners who work in data matching and related areas. It is assumed the reader has some expertise in algorithms and data structures, and database technologies. Most chapters of this book end with a section that provides pointers to further background and research material, which will allow the interested reader to cover gaps in their knowledge and explore a specific topic in more depth.

This book provides the reader with a broad range of data matching concepts and techniques, touching on all aspects of the data matching process. A wide range of research in data matching is covered, and critical comparisons between state-of-the-art approaches are provided. This book can thus help researchers from related fields (such as databases, data mining, machine learning, knowledge engineering, information retrieval, information systems, or health informatics), as well as students who are interested to enter this field of research, to become familiar with recent research developments and identify open research challenges in data matching. Each of the Chaps. 3–9 contain a section that discusses open research topics.

This book can help practitioners to better understand the current state-of-the-art in data matching techniques and concepts. Given that in many application domains it is not feasible to simply use or implement an existing off-the-shelf data matching system without substantial adaption and customisation, it is crucial for practitioners to understand the internal workings and limitations of such systems. Practical considerations are discussed in Chaps. 3–8 for each of the major steps of the data matching process.

The technical level of this book also makes it accessible to students taking advanced undergraduate and graduate level courses on data matching or data quality. While such courses are currently rare, with the ongoing challenges that the areas of data quality and data integration pose in many organizations in both the public and private sectors, there is a demand worldwide for graduates with skills and expertise in these areas. It is hoped that this book can help to address this demand.

Acknowledgments

I would like to start by thanking Tim Churches from the New South Wales Department of Health and Sax Institute, for highlighting in 2001 to me and my

colleagues at the Australian National University that the area of data matching can provide exciting research opportunities, and for supporting our research through funding over several years. Without Tim, much of the outcomes we have accomplished over the past decade, such as the FEBRL data matching system, would not have been possible. Thanks goes also to Ross Gayler and Veda Advantage, David Hawking and Funnelback Pty. Ltd., and Fujitsu Laboratories (Japan). Without their support we would not have been able to continue our research in this area. I also like to acknowledge the funding we received for our research from the Australian Research Council (ARC) under two Linkage Projects (LP0453463 and LP100200079), and from the Australian Partnership for Advanced Computing (APAC).

Along the way, I received advice from experienced data matching practitioners, including William Winkler and John Bass, who emphasized the gap between data matching research and its practical application in the real world. A big thanks goes also to all my students who contributed to our research efforts over the years: Justin Xi Zhu, Puthick Hok, Daniel Belacic, Yinghua Zheng, Xiaoyu Huang, Agus Pudjijono, Irwan Krisna, Karl Goiser, Dinusha Vatsalan, and Zhichun (Sally) Fu.

Large portions of this book were written while I was on sabbatical in 2011, and I would like to thank Henry Gardner, Director Research School of Computer Science at the Australian National University, for facilitating this relief from my normal academic duties. My colleagues Paul Thomas and Richard Jones have provided valuable feedback on early versions of this book, and I would like to thank them for their efforts. Insightful comments by William Winkler, Warwick Graco, and Vassilios Verykios helped to clarify certain aspects of the manuscript.

The list of research challenges and directions provided in Sect. 9.6 was compiled with contributions from Brad Malin, Vassilios Verykios, Hector Garcia-Molina, Steven (Euijong) Whang, Warwick Graco, and William Winkler (who gave the striking comment that "if one goes back 50 + years, these five issues were present" regarding the major challenges of data matching from the perspective of an experienced practitioner).

I would also like to thank the two anonymous reviewers who provided valuable detailed feedback and helpful suggestions. The task of proof-reading of the final manuscript was made easier through the help of my colleagues and students Paul Thomas, Qing Wang, Huizhi (Elly) Liang, Banda Ramadan, Dinusha Vatsalan, Zhichun (Sally) Fu, Felicity Splatt, and Brett Romero, who all detected the small hidden mistakes I had missed.

I also like to thank the editors of this book series, Mike Carey and Stefano Ceri, and to Ralf Gestner from Springer, who all supported this book project right from the start.

And finally, last but not least, a very big thanks goes to Gail for her love, encouragement and understanding.

Canberra, 29 April 2012 Peter Christen

Contents