Peter Christen

Data Matching

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

Springer

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 naïve 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 value-specific 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 methods have 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.

William E. Winkler
U.S. Census Bureau
Suitland, MD, USA

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-the-art 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.

Organisation

This book consists of ten 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 Chapters 3 to 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 chapters 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 Chapter 4 is the different indexing (also known as blocking) techniques that are aimed at reducing the quadratic complexity of the naïve 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 Chapter 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 Chapter 6. Both supervised and unsupervised classification techniques, and pair-wise and collective techniques are presented. Finally, Chapter 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 Chapter 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 Chapter 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 245 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 Chapters 3 to 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 Chapters 3 to 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 organisations 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.

Acknowledgements

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 emphasised 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 Section 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” with regard to 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, *Peter Christen*
29 April 2012

Contents

Part II Steps of the Data Matching Process

References

1. Adly, N.: Efficient record linkage using a double embedding scheme. In: DMIN, pp. 274–281. Las Vegas (2009)
2. Aggarwal, C.C.: Managing and Mining Uncertain Data, *Advances in Database Systems*, vol. 35. Springer (2009)
3. Aggarwal, C.C., Yu, P.S.: The IGrid index: Reversing the dimensionality curse for similarity indexing in high dimensional space. In: ACM SIGKDD, pp. 119–129. Boston (2000)
4. Aggarwal, C.C., Yu, P.S.: Privacy-preserving data mining: models and algorithms, *Advances in Database Systems*, vol. 34. Springer (2008)
5. Agichtein, E., Ganti, V.: Mining reference tables for automatic text segmentation. In: ACM SIGKDD, pp. 20–29. Seattle (2004)
6. Agrawal, R., Evfimievski, A., Srikant, R.: Information sharing across private databases. In: ACM SIGMOD, pp. 86–97. San Diego (2003)
7. Aizawa, A., Oyama, K.: A fast linkage detection scheme for multi-source information integration. In: WIRI, pp. 30–39. Tokyo (2005)
8. Al-Lawati, A., Lee, D., McDaniel, P.: Blocking-aware private record linkage. In: International Workshop on Information Quality in Information Systems, pp. 59–68 (2005)
9. Alvarez, R., Jonas, J., Winkler, W., Wright, R.: Interstate voter registration database matching: the Oregon-Washington 2008 pilot project. In: Workshop on Trustworthy Elections, pp. 17–17. USENIX Association (2009)
10. Anderson, K., Durbin, E., Salinger, M.: Identity theft. The Journal of Economic Perspectives **22**(2), 171–192 (2008)
11. Arasu, A., Götz, M., Kaushik, R.: On active learning of record matching packages. In: ACM SIGMOD, pp. 783–794. Indianapolis (2010)
12. Arasu, A., Kaushik, R.: A grammar-based entity representation framework for data cleaning. In: ACM SIGMOD, pp. 233–244. Providence, Rhode Island (2009)
13. Armstrong, M.P., Ruggles, A.J.: Geographic information technologies and personal privacy. Cartographica: The International Journal for Geographic Information and Geovisualization **40**(4), 63–73 (2005)
14. Atallah, M., Kerschbaum, F., Du, W.: Secure and private sequence comparisons. In: Workshop on Privacy in the Electronic Society, pp. 39–44. ACM (2003)
15. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. Addison-Wesley Longman Publishing Co., Boston (1999)
16. Baldwin, J., Acheson, E., Graham, W.: Textbook of medical record linkage. Oxford University Press (1987)
17. Barone, D., Maurino, A., Stella, F., Batini, C.: A privacy-preserving framework for accuracy and completeness quality assessment. Emerging Paradigms in Informatics, Systems and Communication p. 83 (2009)
18. Bartolini, I., Ciaccia, P., Patella, M.: String matching with metric trees using an approximate distance. In: String Processing and Information Retrieval, LNCS 2476, pp. 271–283. Lisbon, Portugal (2002)
19. Batini, C., Scannapieco, M.: Data quality: Concepts, methodologies and techniques. Data-Centric Systems and Applications. Springer (2006)
20. Baxter, R., Christen, P., Churches, T.: A comparison of fast blocking methods for record linkage. In: ACM SIGKDD Workshop on Data Cleaning, Record Linkage and Object Consolidation, pp. 25–27. Washington DC (2003)
21. Bayardo, R., Ma, Y., Srikant, R.: Scaling up all pairs similarity search. In: WWW, pp. 131–140. Banff, Canada (2007)
22. Behm, A., Ji, S., Li, C., Lu, J.: Space-constrained gram-based indexing for efficient approximate string search. In: IEEE ICDE, pp. 604–615. Shanghai (2009)
23. Belin, T., Rubin, D.: A method for calibrating false-match rates in record linkage. Journal of the American Statistical Association pp. 694–707 (1995)

24. Bellahsene, Z., Bonifati, A., Rahm, E.: Schema Matching and Mapping. Data-Centric Systems and Applications. Springer (2011)
25. Benjelloun, O., Garcia-Molina, H., Gong, H., Kawai, H., Larson, T., Menestrina, D., Thavisomboon, S.: D-Swoosh: A family of algorithms for generic, distributed entity resolution. In: International Conference on Distributed Computing Systems, pp. 37–37 (2007)
26. Benjelloun, O., Garcia-Molina, H., Menestrina, D., Su, Q., Whang, S., Widom, J.: Swoosh: a generic approach to entity resolution. The VLDB Journal **18**(1), 255–276 (2009)
27. Bergroth, L., Hakonen, H., Raita, T.: A survey of longest common subsequence algorithms. In: String Processing and Information Retrieval, pp. 39–48. A Curuna, Spain (2000)
28. Bernecker, T., Kriegel, H.P., Mamoulis, N., Renz, M., Zuefle, A.: Scalable probabilistic similarity ranking in uncertain databases. IEEE Transactions on Knowledge and Data Engineering **22**(9), 1234–1246 (2010)
29. Bertolazzi P De Santis L, S.M.: Automated record matching in cooperative information systems. In: Proceedings of the international workshop on data quality in cooperative information systems. Siena, Italy (2003)
30. Bertsekas, D.P.: Auction algorithms for network flow problems: A tutorial introduction. Computational Optimization and Applications **1**, 7–66 (1992)
31. Bhattacharya, I., Getoor, L.: Collective entity resolution in relational data. ACM Transactions on Knowledge Discovery from Data **1**(1) (2007)
32. Bhattacharya, I., Getoor, L.: Query-time entity resolution. Journal of Artificial Intelligence Research **30**, 621–657 (2007)
33. Bilenko, M., Basu, S., Sahami, M.: Adaptive product normalization: Using online learning for record linkage in comparison shopping. In: IEEE ICDM, pp. 58–65. Houston (2005)
34. Bilenko, M., Kamath, B., Mooney, R.J.: Adaptive blocking: Learning to scale up record linkage. In: IEEE ICDM, pp. 87–96. Hong Kong (2006)
35. Bilenko, M., Mooney, R.J.: Adaptive duplicate detection using learnable string similarity measures. In: ACM SIGKDD, pp. 39–48. Washington DC (2003)
36. Bilgic, M., Licamele, L., Getoor, L., Shneiderman, B.: D-dupe: An interactive tool for entity resolution in social networks. In: IEEE Symposium on Visual Analytics Science and Technology, pp. 43–50 (2006)
37. Blakely T, S.C.: Probabilistic record linkage and a method to calculate the positive predictive value. International Journal of Epidemiology **31:6**, 1246–1252 (2002)
38. Bleiholder, J., Naumann, F.: Data fusion. ACM Computing Surveys **41**(1), 1–41 (2008)
39. Bloom, B.: Space/time trade-offs in hash coding with allowable errors. Communications of the ACM **13**(7), 422–426 (1970)
40. Borgman, C.L., Siegfried, S.L.: Getty's synoname^{TM} and its cousins: A survey of applications of personal name-matching algorithms. Journal of the American Society for Information Science **43**(7), 459–476 (1992)
41. Borkar, V., Deshmukh, K., Sarawagi, S.: Automatic segmentation of text into structured records. ACM SIGMOD Record **30**(2), 175–186 (2001)
42. Breiman, L., Freidman, J., Olshen, R., Stone, C.: Classification and regression trees. Chapman and Hall/CRC (1984)
43. Broder, A., Carmel, D., Herscovici, M., Soffer, A., Zien, J.: Efficient query evaluation using a two-level retrieval process. In: ACM CIKM, pp. 426–434. New Orleans (2003)
44. Brook, E., Rosman, D., Holman, C.: Public good through data linkage: measuring research outputs from the Western Australian data linkage system. Australian and New Zealand journal of public health **32**(1), 19–23 (2008)
45. Brownstein, J.S., Cassa, C., Kohane, I.S., Mandl, K.D.: Reverse geocoding: Concerns about patient confidentiality in the display of geospatial health data. In: AMIA Annual Symposium Proceedings, p. 905. American Medical Informatics Association (2005)
46. Brownstein, J.S., Cassa, C., Mandl, K.D.: No place to hide–reverse identification of patients from published maps. New England Journal of Medicine **355**(16), 1741–1742 (2006)
47. Campbell, K., Deck, D., Krupski, A.: Record linkage software in the public domain: a comparison of Link Plus, The Link King, and a basic deterministic algorithm. Health Informatics Journal **14**(1), 5 (2008)

48. Cayo, M.R., Talbot, T.O.: Positional error in automated geocoding of residential addresses. International Journal of Health Geographics **2**(10) (2003)
49. Cebrián, M., Alfonseca, M., Ortega, A.: Common pitfalls using the normalized compression distance: What to watch out for in a compressor. Communications in Information and Systems **5**(4), 367–384 (2005)
50. Chambers, R.: Regression analysis of probability-linked data. Official Statistics Research Series **4** (2008)
51. Chan, Y., Talburt, J., Talley, T.: Data Engineering. Springer (2010)
52. Chaudhuri, S., Ganti, V., Motwani, R.: Robust identification of fuzzy duplicates. In: IEEE ICDE, pp. 865–876. Tokyo (2005)
53. Chaytor, R., Brown, E., Wareham, T.: Privacy advisors for personal information management. In: SIGIR Workshop on Personal Information Management, pp. 28–31. Seattle, Washington (2006)
54. Chen, H., Chung, W., Xu, J., Wang, G., Qin, Y., Chau, M.: Crime data mining: a general framework and some examples. IEEE Computer **37**(4), 50–56 (2004)
55. Chor, B., Kushilevitz, E., Goldreich, O., Sudan, M.: Private information retrieval. Journal of the ACM (JACM) **45**(6), 965–981 (1998)
56. Christen, P.: Probabilistic data generation for deduplication and data linkage. In: IDEAL, Springer LNCS, vol. 3578, pp. 109–116. Brisbane (2005)
57. Christen, P.: A comparison of personal name matching: Techniques and practical issues. In: Workshop on Mining Complex Data, held at IEEE ICDM. Hong Kong (2006)
58. Christen, P.: Privacy-preserving data linkage and geocoding: Current approaches and research directions. In: Workshop on Privacy Aspects of Data Mining, held at IEEE ICDM. Hong Kong (2006)
59. Christen, P.: Automatic record linkage using seeded nearest neighbour and support vector machine classification. In: ACM SIGKDD, pp. 151–159. Las Vegas (2008)
60. Christen, P.: Automatic training example selection for scalable unsupervised record linkage. In: PAKDD, Springer LNAI, vol. 5012, pp. 511–518. Osaka (2008)
61. Christen, P.: Febrl: An open source data cleaning, deduplication and record linkage system with a graphical user interface. In: ACM SIGKDD, pp. 1065–1068. Las Vegas (2008)
62. Christen, P.: Development and user experiences of an open source data cleaning, deduplication and record linkage system. SIGKDD Explorations **11**(1), 39–48 (2009)
63. Christen, P.: Geocode matching and privacy preservation. In: Workshop on Privacy, Security, and Trust in KDD, pp. 7–24. Springer (2009)
64. Christen, P.: A survey of indexing techniques for scalable record linkage and deduplication. IEEE Transactions on Knowledge and Data Engineering **X**(Y) (2011)
65. Christen, P., Belacic, D.: Automated probabilistic address standardisation and verification. In: AusDM, pp. 53–67. Sydney (2005)
66. Christen, p., Churches, T., Hegland, M.: Febrl – A parallel open source data linkage system. In: PAKDD, Springer LNAI, vol. 3056, pp. 638–647. Sydney (2004)
67. Christen, P., Churches, T., Willmore, A.: A probabilistic geocoding system based on a national address file. In: AusDM. Cairns (2004)
68. Christen, P., Churches, T., Zhu, J.: Probabilistic name and address cleaning and standardization. In: Australasian Data Mining Workshop. Canberra (2002)
69. Christen, P., Gayler, R.: Towards scalable real-time entity resolution using a similarity-aware inverted index approach. In: AusDM, CRPIT, vol. 87, pp. 51–60. Glenelg, Australia (2008)
70. Christen, P., Gayler, R., Hawking, D.: Similarity-aware indexing for real-time entity resolution. In: ACM CIKM, pp. 1565–1568. Hong Kong (2009)
71. Christen, P., Goiser, K.: Quality and complexity measures for data linkage and deduplication. In: F. Guillet, H. Hamilton (eds.) Quality Measures in Data Mining, *Studies in Computational Intelligence*, vol. 43, pp. 127–151. Springer (2007)
72. Christen, P., Pudjijono, A.: Accurate synthetic generation of realistic personal information. In: PAKDD, Springer LNAI, vol. 5476, pp. 507–514. Bangkok, Thailand (2009)

73. Churches, T.: A proposed architecture and method of operation for improving the protection of privacy and confidentiality in disease registers. BioMed Central Medical Research Methodology **3**(1) (2003)
74. Churches, T., Christen, P.: Blind data linkage using n-gram similarity comparisons. In: PAKDD, Springer LNAI, vol. 3056, pp. 121–126. Sydney (2004)
75. Churches, T., Christen, P.: Some methods for blindfolded record linkage. BioMed Central Medical Informatics and Decision Making **4**(9) (2004)
76. Churches, T., Christen, P., Lim, K., Zhu, J.X.: Preparation of name and address data for record linkage using hidden Markov models. BioMed Central Medical Informatics and Decision Making **2**(9) (2002)
77. Cilibrasi, R., Vitányi, P.M.: Clustering by compression. IEEE Transactions on Information Theory **51**(4), 1523–1545 (2005)
78. Clark, D.E.: Practical introduction to record linkage for injury research. Injury Prevention **10**, 186–191 (2004)
79. Clifton, C., Kantarcioglu, M., Doan, A., Schadow, G., Vaidya, J., Elmagarmid, A., Suciu, D.: Privacy-preserving data integration and sharing. In: ACM SIGMOD workshop on Research issues in Data Mining and Knowledge Discovery, pp. 19–26 (2004)
80. Cochinwala, M., Kurien, V., Lalk, G., Shasha, D.: Efficient data reconciliation. Information Sciences **137**(1–4), 1–15 (2001)
81. Cohen, W.: The WHIRL approach to data integration. IEEE Intelligent Systems **13**(3), 20–24 (1998)
82. Cohen, W.: Data integration using similarity joins and a word-based information representation language. ACM Transactions on Information Systems **18**(3), 288–321 (2000)
83. Cohen, W.W.: Integration of heterogeneous databases without common domains using queries based on textual similarity. In: ACM SIGMOD, pp. 201–212. Seattle (1998)
84. Cohen, W.W., Ravikumar, P., Fienberg, S.: A comparison of string distance metrics for name-matching tasks. In: Workshop on Information Integration on the Web, held at IJCAI, pp. 73–78. Acapulco (2003)
85. Cohen, W.W., Richman, J.: Learning to match and cluster large high-dimensional data sets for data integration. In: ACM SIGKDD, pp. 475–480. Edmonton (2002)
86. Conn, L., Bishop, G.: Exploring methods for creating a longitudinal census dataset. Tech. Rep. 1352.0.55.076, Australian Bureau of Statistics, Canberra (2005)
87. Curtis, A.J., Mills, J.W., Leitner, M.: Spatial confidentiality and GIS: Re-engineering mortality locations from published maps about Hurricane Katrina. International Journal of Health Geographics **5**(1), 44–56 (2006)
88. Dal Bianco, G., Galante, R., Heuser, C.: A fast approach for parallel deduplication on multicore processors. In: ACM Symposium on Applied Computing, pp. 1027–1032 (2011)
89. Damerau, F.J.: A technique for computer detection and correction of spelling errors. Communications of the ACM **7**(3), 171–176 (1964)
90. Day, C.: Record linkage i: evaluation of commercially available record linkage software for use in NASS. Tech. Rep. STB Research Report STB-95-02, National Agricultural Statistics Service, Washington DC (1995)
91. Dey, D., Mookerjee, V., Liu, D.: Efficient techniques for online record linkage. IEEE Transactions on Knowledge and Data Engineering **23**(3), 373–387 (2010)
92. Domingo-Ferrer, J., Torra, V.: Disclosure risk assessment in statistical microdata protection via advanced record linkage. Statistics and Computing **13**(4), 343–354 (2003)
93. Dong, X., Halevy, A., Madhavan, J.: Reference reconciliation in complex information spaces. In: ACM SIGMOD, pp. 85–96. Baltimore (2005)
94. Draisbach, U., Naumann, F.: A comparison and generalization of blocking and windowing algorithms for duplicate detection. In: Workshop on Quality in Databases, held at VLDB. Lyon (2009)
95. Draisbach, U., Naumann, F.: Dude: The duplicate detection toolkit. In: Workshop on Quality in Databases, held at VLDB. Singapore (2010)

96. Du, W., Atallah, M., Kerschbaum, F.: Protocols for secure remote database access with approximate matching. In: First ACM Workshop on Security and Privacy in E-Commerce (2000)
97. Dunn, H.: Record linkage. American Journal of Public Health **36**(12), 1412 (1946)
98. Durham, E., Xue, Y., Kantarcioglu, M., Malin, B.: Private medical record linkage with approximate matching. In: AMIA Annual Symposium Proceedings, p. 182. American Medical Informatics Association (2010)
99. Durham, E., Xue, Y., Kantarcioglu, M., Malin, B.: Quantifying the correctness, computational complexity, and security of privacy-preserving string comparators for record linkage. Information Fusion **In Press** (2011)
100. Durham, E.A.: A framework for accurate, efficient private record linkage. Ph.D. thesis, Faculty of the Graduate School of Vanderbilt University, Nashville, TN (2012)
101. Dwork, C.: Differential privacy. Automata, languages and programming pp. 1–12 (2006)
102. Elfeky, M.G., Verykios, V.S., Elmagarmid, A.K.: TAILOR: A record linkage toolbox. In: IEEE ICDE, pp. 17–28. San Jose (2002)
103. Elmagarmid, A.K., Ipeirotis, P.G., Verykios, V.S.: Duplicate record detection: A survey. IEEE Transactions on Knowledge and Data Engineering **19**(1), 1–16 (2007)
104. Fagin, R., Naor, M., Winkler, P.: Comparing information without leaking it. Communications of the ACM **39**(5), 77–85 (1996)
105. Faloutsos, C., Lin, K.I.: Fastmap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets. In: ACM SIGMOD, pp. 163–174. San Jose (1995)
106. Fawcett T: ROC Graphs: Notes and practical considerations for researchers. Tech. Rep. HPL-2003-4, HP Laboratories, Palo Alto (2004)
107. Fellegi, I.P., Holt, D.: A systematic approach to automatic edit and imputation. Journal of the American Statistical Association pp. 17–35 (1976)
108. Fellegi, I.P., Sunter, A.B.: A theory for record linkage. Journal of the American Statistical Association **64**(328), 1183–1210 (1969)
109. Fienberg, S.: Homeland insecurity: Datamining, terrorism detection, and confidentiality. Bull. Internat. Stat. Inst (2005)
110. Fienberg, S.: Privacy and confidentiality in an e-commerce world: Data mining, data warehousing, matching and disclosure limitation. Statistical Science **21**(2), 143–154 (2006)
111. Fogel, R.: New sources and new techniques for the study of secular trends in nutritional status, health, mortality, and the process of aging. NBER Historical Working Papers (1993)
112. Fortini, M., Liseo, B., Nuccitelli, A., Scanu, M.: On Bayesian record linkage. Research in Official Statistics **4**(1), 185–198 (2001)
113. Friedman, C., Sideli, R.: Tolerating spelling errors during patient validation. Computers and Biomedical Research **25**, 486–509 (1992)
114. Fu, Z., Christen, P., Boot, M.: Automatic cleaning and linking of historical census data using household information. In: Workshop on Domain Driven Data Mining, held at IEEE ICDM. Vancouver (2011)
115. Fu, Z., Christen, P., Boot, M.: A supervised learning and group linking method for historical census household linkage. In: AusDM, CRPIT, vol. 125. Ballarat, Australia (2011)
116. Fu, Z., Zhou, J., Christen, P., Boot, M.: Multiple instance learning for group record linkage. In: PAKDD, Springer LNAI. Kuala Lumpur, Malaysia (2012)
117. Galhardas, H., Florescu, D., Shasha, D., Simon, E.: An extensible framework for data cleaning. In: IEEE ICDE. San Diego (2000)
118. Gill, L.: OX-LINK: The Oxford medical record linkage system. In: Proc. Intl Record Linkage Workshop and Exposition, pp. 15–33. Arlington, Virginia (1997)
119. Gill, L.: Methods for automatic record matching and linking and their use in national statistics. Tech. Rep. Methodology Series, no. 25, National Statistics, London (2001)
120. Giunchiglia, F., Yatskevich, M., Shvaiko, P.: Semantic matching: Algorithms and implementation. Journal on Data Semantics IX pp. 1–38 (2007)

121. Glasson, E., De Klerk, N., Bass, A., Rosman, D., Palmer, L., Holman, C.: Cohort profile: the Western Australian family connections genealogical project. International Journal of epidemiology **37**(1), 30–35 (2008)
122. Gliklich, R., Dreyer, N. (eds.): Registries for Evaluating Patient Outcomes: A Users Guide. No.10-EHC049. AHRQ Publication (2010)
123. Goldreich, O.: Secure multi-party computation. Tech. rep., Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Israel (2002)
124. Gomatam, S., Carter, R., Ariet, M., Mitchell, G.: An empirical comparison of record linkage procedures. Statistics in Medicine **21**(10), 1485–1496 (2002)
125. Gong, R., Chan, T.K.: Syllable alignment: A novel model for phonetic string search. IEICE Transactions on Information and Systems **E89-D**(1), 332–339 (2006)
126. Grama, A., Karypis, G., Kumar, V., Gupta, A.: Introduction to parallel computing, 2 edn. Addison-Wesley Longman Publishing Co., Inc. (2003)
127. Gravano, L., Ipeirotis, P.G., Jagadish, H.V., Koudas, N., Muthukrishnan, S., Srivastava, D.: Approximate string joins in a database (almost) for free. In: VLDB, pp. 491–500. Roma (2001)
128. Gu, L., Baxter, R.: Adaptive filtering for efficient record linkage. In: SIAM international conference on data mining. Orlando, Florida (2004)
129. Gu, L., Baxter, R.: Decision models for record linkage. In: Selected Papers from AusDM, Springer LNCS 3755, pp. 146–160 (2006)
130. Guo, H., Zhu, H., Guo, Z., Zhang, X., Su, Z.: Address standardization with latent semantic association. In: ACM SIGKDD, pp. 1155–1164. Paris (2009)
131. Hajishirzi, H., Yih, W., Kolcz, A.: Adaptive near-duplicate detection via similarity learning. In: ACM SIGIR, pp. 419–426. Geneva, Switzerland (2010)
132. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.: The WEKA data mining software: an update. ACM SIGKDD Explorations **11**(1), 10–18 (2009)
133. Hall, P.A., Dowling, G.R.: Approximate string matching. ACM Computing Surveys **12**(4), 381–402 (1980)
134. Hall, R., Fienberg, S.: Privacy-preserving record linkage. In: Privacy in Statistical Databases, Springer LNCS 6344, pp. 269–283. Corfu, Greece (2010)
135. Han, J., Kamber, M.: Data mining: concepts and techniques, 2 edn. Morgan Kaufmann (2006)
136. Hand, D.: Classifier technology and the illusion of progress. Statistical Science **21**(1), 1–14 (2006)
137. Hassanzadeh, O., Miller, R.: Creating probabilistic databases from duplicated data. The VLDB Journal **18**(5), 1141–1166 (2009)
138. Heckerman, D.: Bayesian networks for data mining. Data mining and knowledge discovery **1**(1), 79–119 (1997)
139. Henzinger, M.: Finding near-duplicate web pages: a large-scale evaluation of algorithms. In: ACM SIGIR, pp. 284–291. Seattle (2006)
140. Hernandez, M.A., Stolfo, S.J.: The merge/purge problem for large databases. In: ACM SIGMOD, pp. 127–138. San Jose (1995)
141. Hernandez, M.A., Stolfo, S.J.: Real-world data is dirty: Data cleansing and the merge/purge problem. Data Mining and Knowledge Discovery **2**(1), 9–37 (1998)
142. Herschel, M., Naumann, F., Szott, S., Taubert, M.: Scalable iterative graph duplicate detection. IEEE Transactions on Knowledge and Data Engineering **X**(Y) (2011)
143. Herzog, T., Scheuren, F., Winkler, W.: Data quality and record linkage techniques. Springer Verlag (2007)
144. Hirsch, J.: An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences of the United States of America **102**(46), 16,569–16,572 (2005)
145. Holmes, D., McCabe, C.M.: Improving precision and recall for Soundex retrieval. In: Proceedings of the IEEE International Conference on Information Technology – Coding and Computing. Las Vegas (2002)

146. Inan, A., Kantarcioglu, M., Bertino, E., Scannapieco, M.: A hybrid approach to private record linkage. In: IEEE ICDE, pp. 496–505 (2008)
147. Inan, A., Kantarcioglu, M., Ghinita, G., Bertino, E.: Private record matching using differential privacy. In: International Conference on Extending Database Technology, pp. 123–134 (2010)
148. Ioannou, E., Nejdl, W., Niederée, C., Velegrakis, Y.: On-the-fly entity-aware query processing in the presence of linkage. Proceedings of the VLDB Endowment **3**(1) (2010)
149. Jaro, M.A.: Advances in record-linkage methodology a applied to matching the 1985 Census of Tampa, Florida. Journal of the American Statistical Association **84**, 414–420 (1989)
150. Jentzsch, A., Isele, R., Bizer, C.: Silk–generating RDF links while publishing or consuming linked data. In: Poster at the International Semantic Web Conference. Shanghai (2010)
151. Jin, L., Li, C., Mehrotra, S.: Efficient record linkage in large data sets. In: DASFAA, pp. 137–146. Tokyo (2003)
152. Jokinen, P., Tarhio, J., Ukkonen, E.: A comparison of approximate string matching algorithms. Software – Practice and Experience **26**(12), 1439–1458 (1996)
153. Jonas, J., Harper, J.: Effective counterterrorism and the limited role of predictive data mining. Policy Analysis (584) (2006)
154. Jurczyk, P., Lu, J., Xiong, L., Cragan, J., Correa, A.: FRIL: A tool for comparative record linkage. In: AMIA Annual Symposium Proceedings, p. 440. American Medical Informatics Association (2008)
155. Kalashnikov, D., Mehrotra, S.: Domain-independent data cleaning via analysis of entity-relationship graph. ACM Transactions on Database Systems **31**(2), 716–767 (2006)
156. Kang, H., Getoor, L., Shneiderman, B., Bilgic, M., Licamele, L.: Interactive entity resolution in relational data: A visual analytic tool and its evaluation. IEEE Transactions on Visualization and Computer Graphics **14**(5), 999–1014 (2008)
157. Karakasidis, A., Verykios, V.: Privacy preserving record linkage using phonetic codes. In: Fourth Balkan Conference in Informatics, pp. 101–106. IEEE (2009)
158. Karakasidis, A., Verykios, V.: Advances in privacy preserving record linkage. In: E-activity and Innovative Technology, Advances in Applied Intelligence Technologies Book Series, pp. 22–34. IGI Global (2010)
159. Karakasidis, A., Verykios, V., Christen, P.: Fake injection strategies for private phonetic matching. In: International Workshop on Data Privacy Management. Leuven, Belgium (2011)
160. Kawai, H., Garcia-Molina, H., Benjelloun, O., Menestrina, D., Whang, E., Gong, H.: P-Swoosh: Parallel algorithm for generic entity resolution. Tech. Rep. 2006-19, Department of Computer Science, Stanford University (2006)
161. Kelman, C.W., Bass, J., Holman, D.: Research use of linked health data – A best practice protocol. Aust NZ Journal of Public Health **26**, 251–255 (2002)
162. Keskustalo, H., Pirkola, A., Visala, K., Leppanen, E., Jarvelin, K.: Non-adjacent digrams improve matching of cross-lingual spelling variants. In: String Processing and Information Retrieval, LNCS 2857, pp. 252–265. Manaus, Brazil (2003)
163. Kim, H., Lee, D.: Parallel linkage. In: ACM CIKM, pp. 283–292. Lisboa, Portugal (2007)
164. Kim, H., Lee, D.: Harra: fast iterative hashed record linkage for large-scale data collections. In: International Conference on Extending Database Technology, pp. 525–536. Lausanne, Switzerland (2010)
165. Kirsten, T., Kolb, L., Hartung, M., Gross, A., Köpcke, H., Rahm, E.: Data partitioning for parallel entity matching. Proceedings of the VLDB Endowment **3**(2) (2010)
166. Klenk, S., Thom, D., Heidemann, G.: The normalized compression distance as a distance measure in entity identification. Advances in Data Mining. Applications and Theoretical Aspects pp. 325–337 (2009)
167. Kolb, L., Thor, A., Rahm, E.: Multi-pass sorted neighborhood blocking with Map-Reduce. Computer Science-Research and Development pp. 1–19 (2011)
168. Köpcke, H., Rahm, E.: Frameworks for entity matching: A comparison. Data and Knowledge Engineering **69**(2), 197–210 (2010)

169. Köpcke, H., Thor, A., Rahm, E.: Evaluation of entity resolution approaches on real-world match problems. Proceedings of the VLDB Endowment **3**(1-2), 484–493 (2010)
170. Koudas, N., Marathe, A., Srivastava, D.: Flexible string matching against large databases in practice. In: VLDB, pp. 1086–1094. Toronto (2004)
171. Krouse, W., Elias, B.: Terrorist Watchlist Checks and Air Passenger Prescreening. RL33645. Congressional Research Service (2009). CRS Report for Congress
172. Kukich, K.: Techniques for automatically correcting words in text. ACM Computing Surveys **24**(4), 377–439 (1992)
173. Kuzu, M., Kantarcioglu, M., Durham, E., Malin, B.: A constraint satisfaction cryptanalysis of Bloom filters in private record linkage. In: Privacy Enhancing Technologies, pp. 226–245. Springer (2011)
174. Lahiri, P., Larsen, M.: Regression analysis with linked data. Journal of the American statistical association **100**(469), 222–230 (2005)
175. Lait, A., Randell, B.: An assessment of name matching algorithms. Tech. rep., Department of Computer Science, University of Newcastle upon Tyne (1993)
176. Lee, D., Kang, J., Mitra, P., Giles, C.L., On, B.W.: Are your citations clean? Commununications of the ACM **50**, 33–38 (2007)
177. Lee, Y., Pipino, L., Funk, J., Wang, R.: Journey to data quality. The MIT Press (2009)
178. Lenzerini, M.: Data integration: A theoretical perspective. In: ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, pp. 233–246. Madison (2002)
179. Li, P., Dong, X., Maurino, A., Srivastava, D.: Linking temporal records. Proceedings of the VLDB Endowment **4**(11) (2011)
180. Malin, B.: K-unlinkability: A privacy protection model for distributed data. Data and Knowledge Engineering **64**(1), 294–311 (2008)
181. Malin, B., Airoldi, E., Carley, K.: A network analysis model for disambiguation of names in lists. Computational and Mathematical Organization Theory **11**(2), 119–139 (2005)
182. Malin, B., Karp, D., Scheuermann, R.: Technical and policy approaches to balancing patient privacy and data sharing in clinical and translational research. Journal of investigative medicine: the official publication of the American Federation for Clinical Research **58**(1), 11 (2010)
183. Manghi, P., Mikulicic, M.: PACE: A general-purpose tool for authority control. Metadata and Semantic Research pp. 80–92 (2011)
184. McCallum, A., Nigam, K., Rennie, J., Seymore, K.: Automating the construction of Internet portals with machine learning. Information Retrieval **3**(2), 127–163 (2000)
185. McCallum, A., Nigam, K., Ungar, L.H.: Efficient clustering of high-dimensional data sets with application to reference matching. In: ACM SIGKDD, pp. 169–178. Boston (2000)
186. Menestrina, D., Benjelloun, O., Garcia-Molina, H.: Generic entity resolution with data confidences. In: First International VLDB Workshop on Clean Databases. Seoul, South Korea (2006)
187. Menestrina, D., Whang, S., Garcia-Molina, H.: Evaluating entity resolution results. Proceedings of the VLDB Endowment **3**(1–2), 208–219 (2010)
188. Michelson, M., Knoblock, C.A.: Learning blocking schemes for record linkage. In: AAAI. Boston (2006)
189. Mitchell, T.M.: Machine Learning. McGraw Hill (1997)
190. Monge, A.E.: Matching algorithms within a duplicate detection system. IEEE Data Engineering Bulletin **23**(4), 14–20 (2000)
191. Monge, A.E., Elkan, C.P.: The field-matching problem: Algorithm and applications. In: ACM SIGKDD, pp. 267–270. Portland (1996)
192. Moreau, E., Yvon, F., Cappé, O.: Robust similarity measures for named entities matching. In: 22nd International Conference on Computational Linguistics-Volume 1, pp. 593–600. Association for Computational Linguistics (2008)
193. Moustakides, G.V., Verykios, V.S.: Optimal stopping: A record linkage approach. Journal Data and Information Quality **1**, 9:1–9:34 (2009)
194. Narayanan, A., Shmatikov, V.: Myths and fallacies of personally identifiable information. Communications of the ACM **53**(6), 24–26 (2010)

195. Naumann, F., Herschel, M.: An introduction to duplicate detection, *Synthesis Lectures on Data Management*, vol. 3. Morgan and Claypool Publishers (2010)
196. Navarro, G.: A guided tour to approximate string matching. ACM Computing Surveys **33**(1), 31–88 (2001)
197. Newcombe, H., Kennedy, J.: Record linkage: making maximum use of the discriminating power of identifying information. Communications of the ACM **5**(11), 563–566 (1962)
198. Newcombe, H., Kennedy, J., Axford, S., James, A.: Automatic linkage of vital records. Science **130**(3381), 954–959 (1959)
199. Newcombe, H.B.: Handbook of record linkage: methods for health and statistical studies, administration, and business. Oxford University Press, Inc., New York, NY, USA (1988)
200. Nin, J., Muntes-Mulero, V., Martinez-Bazan, N., Larriba-Pey, J.L.: On the use of semantic blocking techniques for data cleansing and integration. In: IDEAS, pp. 190–198. Banff, Canada (2007)
201. Odell, M., Russell, R.: The soundex coding system. US Patents **1261167** (1918)
202. O'Keefe, C., Yung, M., Gu, L., Baxter, R.: Privacy-preserving data linkage protocols. In: ACM Workshop on Privacy in the Electronic Society, pp. 94–102. Washington DC (2004)
203. Okner, B.: Data matching and merging: An overview. NBER Chapters pp. 49–54 (1974)
204. On, B.W., Elmacioglu, E., Lee, D., Kang, J., Pei, J.: Improving grouped-entity resolution using quasi-cliques. In: IEEE ICDM, pp. 1008–1015 (2006)
205. On, B.W., Koudas, N., Lee, D., Srivastava, D.: Group linkage. In: IEEE ICDE, pp. 496–505. Istanbul (2007)
206. Oscherwitz, T.: Synthetic identity fraud: unseen identity challenge. Bank Security News **3(**7) (2005)
207. Pang, C., Gu, L., Hansen, D., Maeder, A.: Privacy-preserving fuzzy matching using a public reference table. Intelligent Patient Management pp. 71–89 (2009)
208. Patman, F., Thompson, P.: Names: A new frontier in text mining. In: ISI-2003, Springer LNCS 2665, pp. 27–38 (2003)
209. Paull, D.: A geocoded national address file for Australia: The G-NAF what, why, who and when? PSMA Australia Limited, Griffith, ACT, Australia (2003)
210. Pfeifer, U., Poersch, T., Fuhr, N.: Retrieval effectiveness of proper name search methods. Information Processing and Management **32**(6), 667–679 (1996)
211. Philips, L.: The double-metaphone search algorithm. C/C++ User's Journal **18**(6) (2000)
212. Phua, C., Smith-Miles, K., Lee, V., Gayler, R.: Resilient identity crime detection. IEEE Transactions on Knowledge and Data Engineering **24**(3) (2012)
213. Poindexter, J., Popp, R., Sharkey, B.: Total information awareness (TIA). In: IEEE Aerospace Conference, 2003, vol. 6, pp. 2937–2944 (2003)
214. Pollock, J.J., Zamora, A.: Automatic spelling correction in scientific and scholarly text. Communications of the ACM **27**(4), 358–368 (1984)
215. Porter, E.H., Winkler, W.E.: Approximate string comparison and its effect on an advanced record linkage system. Tech. Rep. RR97/02, US Bureau of the Census (1997)
216. Prabhakar, S., Shah, R., Singh, S.: Indexing uncertain data. In: C.C. Aggarwal (ed.) Managing and Mining Uncertain Data, *Advances in Database Systems*, vol. 35, pp. 299–325. Springer (2009)
217. Prasad, K., Faruquie, T., Joshi, S., Chaturvedi, S., Subramaniam, L., Mohania, M.: Data cleansing techniques for large enterprise datasets. In: SRII Global Conference, pp. 135–144. San Jose, USA (2009)
218. Pyle, D.: Data preparation for data mining. Morgan Kaufmann (1999)
219. Quantin, C., Bouzelat, H., Allaert, F., Benhamiche, A., Faivre, J., Dusserre, L.: How to ensure data quality of an epidemiological follow-up: Quality assessment of an anonymous record linkage procedure. International Journal of Medical Informatics **49**(1), 117–122 (1998)
220. Quantin, C., Bouzelat, H., Allaert, F.A., Benhamiche, A.M., Faivre, J., Dusserre, L.: Automatic record hash coding and linkage for epidemiological follow-up data confidentiality. Methods of Information in Medicine **37**(3), 271–277 (1998)
221. Quantin, C., Bouzelat, H., Dusserre, L.: Irreversible encryption method by generation of polynomials. Medical Informatics and the Internet in Medicine **21**(2), 113–121 (1996)

222. Quass, D., Starkey, P.: Record linkage for genealogical databases. In: ACM SIGKDD Workshop on Data Cleaning, Record Linkage and Object Consolidation, pp. 40–42. Washington DC (2003)
223. Rabiner, L.: A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE **77**(2), 257–286 (1989)
224. Rahm, E., Do, H.H.: Data cleaning: Problems and current approaches. IEEE Data Engineering Bulletin **23**(4), 3–13 (2000)
225. Rastogi, V., Dalvi, N., Garofalakis, M.: Large-scale collective entity matching. VLDB Endowment **4**, 208–218 (2011)
226. Ravikumar, P., Cohen, W., Fienberg, S.: A secure protocol for computing string distance metrics. In: Workshop on Privacy and Security Aspects of Data Mining held at IEEE ICDM, pp. 40–46. Brighton, UK (2004)
227. Ruggles, S.: Linking historical censuses: A new approach. History and Computing **14**(1–2), 213–224 (2002)
228. Rushton, G., Armstrong, M., Gittler, J., Greene, B., Pavlik, C., West, M., Zimmerman, D.: Geocoding in cancer research: A review. American Journal of Preventive Medicine **30**(2), S16–S24 (2006)
229. Sadinle, M., Hall, R., Fienberg, S.: Approaches to multiple record linkage. Proceedings of International Statistical Institute (2011)
230. Sarawagi, S.: Information extraction. Foundations and Trends in Databases **1**(3), 261–377 (2008)
231. Sarawagi, S., Bhamidipaty, A.: Interactive deduplication using active learning. In: ACM SIGKDD, pp. 269–278. Edmonton (2002)
232. Sarawagi, S., Kirpal, A.: Efficient set joins on similarity predicates. In: ACM SIGMOD, pp. 754–765. Paris (2004)
233. Sariyar, M., Borg, A.: The RecordLinkage package: Detecting errors in data. The R Journal **2**(2), 61–67 (2010)
234. Scannapieco, M., Figotin, I., Bertino, E., Elmagarmid, A.: Privacy preserving schema and data matching. In: ACM SIGMOD, pp. 653–664 (2007)
235. Scheuren, F., Winkler, W.: Regression analysis of data files that are computer matched. Statistics of income: Turning administrative systems into information systems **1299**(1), 131 (1993)
236. Schewe, K., Wang, Q.: On the decidability and complexity of identity knowledge representation. In: Database Systems for Advanced Applications, Springer LNCS 7238, pp. 288–302. Busan, South Korea (2012)
237. Schneier, B.: Applied cryptography: Protocols, algorithms, and source code in C, 2 edn. John Wiley and Sons, Inc., New York (1996)
238. Schnell, R., Bachteler, T., Bender, S.: A toolbox for record linkage. Austrian Journal of Statistics **33**(1 & 2), 125–133 (2004)
239. Schnell, R., Bachteler, T., Reiher, J.: Privacy-preserving record linkage using Bloom filters. BioMed Central Medical Informatics and Decision Making **9**(1) (2009)
240. Seymore, K., McCallum, A., Rosenfeld, R.: Learning hidden Markov model structure for information extraction. In: AAAI Workshop on Machine Learning for Information Extraction, pp. 37–42 (1999)
241. Smith, M., Newcombe, H.: Methods for computer linkage of hospital admission-separation records into cumulative health histories. Methods of Information in Medicine **14**(3), 118–125 (1975)
242. Smith, M., Newcombe, H.: Accuracies of computer versus manual linkages of routine health records. Methods of Information in Medicine **18**(2), 89–97 (1979)
243. Snae, C.: A comparison and analysis of name matching algorithms. International Journal of Applied Science, Engineering and Technology **4**(1), 252–257 (2007)
244. Song, D., Wagner, D., Perrig, A.: Practical techniques for searches on encrypted data. In: IEEE Symposium on Security and Privacy, pp. 44–55 (2000)
245. Su, W., Wang, J., Lochovsky, F.H.: Record matching over query results from multiple web databases. IEEE Transactions on Knowledge and Data Engineering **22**(4), 578–589 (2009)

246. Summerhayes, R., Holder, P., Beard, J., Morgan, G., Christen, P., Willmore, A., Churches, T.: Automated geocoding of routinely collected health data in New South Wales. New South Wales Public Health Bulletin **17**(4), 33–38 (2006)
247. Sweeney, L.: Computational disclosure control: A primer on data privacy protection. Ph.D. thesis, Massachusetts Institute of Technology, Dept. of Electrical Engineering and Computer Science (2001)
248. Sweeney, L.: K-anonymity: A model for protecting privacy. International Journal of Uncertainty Fuzziness and Knowledge Based Systems **10**(5), 557–570 (2002)
249. Talburt, J.: Entity Resolution and Information Quality. Morgan Kaufmann (2011)
250. Talburt, J.R., Zhou, Y., Shivaiah, S.Y.: SOG: A synthetic occupancy generator to support entity resolution instruction and research. In: International Conference on Information Quality, pp. 91–105. Potsdam, Germany (2009)
251. Technologies, M.: AutoStan and AutoMatch, User's Manuals (1998). Kennebunk, Maine
252. Tejada, S., Knoblock, C.A., Minton, S.: Learning domain-independent string transformation weights for high accuracy object identification. In: ACM SIGKDD, pp. 350–359. Edmonton (2002)
253. Torra, V., Domingo-Ferrer, J.: Record linkage methods for multidatabase data mining. Studies in Fuzziness and Soft Computing **123**, 101–132 (2003)
254. Torra, V., Domingo-Ferrer, J., Torres, A.: Data mining methods for linking data coming from several sources. In: Third Joint UN/ECE-Eurostat Work Session on Statistical Data Confidentiality, Eurostat. Monographs in Official Statistics. Luxembourg (2004)
255. Trepetin, S.: Privacy-preserving string comparisons in record linkage systems: a review. Information Security Journal: A Global Perspective **17**(5), 253–266 (2008)
256. US Federal Geographic Data Committee. Homeland Security and Geographic Information Systems: How GIS and mapping technology can save lives and protect property in post-September 11th America. Public Health GIS News and Information (52), 21–23 (2003)
257. Vaidya, J., Clifton, C., Zhu, M.: Privacy preserving data mining, vol. 19. Springer (2006)
258. Van Berkel, B., De Smedt, K.: Triphone analysis: A combined method for the correction of orthographical and typographical errors. In: Second Conference on Applied Natural Language Processing, pp. 77–83. Austin (1988)
259. Vapnik, V.: The nature of statistical learning theory. Springer (2000)
260. Vatsalan, D., Christen, P., Verykios, V.: An efficient two-party protocol for approximate matching in private record linkage. In: AusDM, CRPIT, vol. 121. Ballarat, Australia (2011)
261. Verykios, V., Elmagarmid, A., Houstis, E.: Automating the approximate record-matching process. Information Sciences **126**(1–4), 83–98 (2000)
262. Verykios, V., Karakasidis, A., Mitrogiannis, V.: Privacy preserving record linkage approaches. Int. J. of Data Mining, Modelling and Management **1**(2), 206–221 (2009)
263. Verykios, V.S., George, M.V., Elfeky, M.G.: A Bayesian decision model for cost optimal record matching. The VLDB Journal **12**(1), 28–40 (2003)
264. Volz, J., Bizer, C., Gaedke, M., Kobilarov, G.: Silk–a link discovery framework for the web of data. In: Second Linked Data on the Web Workshop (2009)
265. de Vries, T., Ke, H., Chawla, S., Christen, P.: Robust record linkage blocking using suffix arrays. In: ACM CIKM, pp. 305–314. Hong Kong (2009)
266. de Vries, T., Ke, H., Chawla, S., Christen, P.: Robust record linkage blocking using suffix arrays and Bloom filters. ACM Transactions on Knowledge Discovery from Data **5**(2) (2011)
267. Wang, G., Chen, H., Atabakhsh, H.: Automatically detecting deceptive criminal identities. Communications of the ACM **47**(3), 70–76 (2004)
268. Wartell, J., McEwen, T.: Privacy in the information age: A guide for sharing crime maps and spatial data. Institute for Law and Justice, NCJ 188739 (2001)
269. Weis, M., Naumann, F.: Detecting duplicate objects in xml documents. In: International Workshop on Information Quality in Information Systems, pp. 10–19. Paris (2004)
270. Weis, M., Naumann, F.: Dogmatix tracks down duplicates in XML. In: ACM SIGMOD, pp. 431–442. Baltimore (2005)
271. Weis, M., Naumann, F., Brosy, F.: A duplicate detection benchmark for XML (and relational) data. In: Workshop on Information Quality for Information Systems (IQIS). Chicago (2006)

272. Weis, M., Naumann, F., Jehle, U., Lufter, J., Schuster, H.: Industry-scale duplicate detection. Proceedings of the VLDB Endowment **1**(2), 1253–1264 (2008)
273. West, D.: Introduction to graph theory, 3 edn. Prentice Hall (2007)
274. Whang, S., Garcia-Molina, H.: Entity resolution with evolving rules. Proceedings of the VLDB Endowment **3**(1-2), 1326–1337 (2010)
275. Whang, S.E., Garcia-Molina, H.: Developments in generic entity resolution. IEEE Data Engineering Bulletin **34**(3), 51–59 (2011)
276. Whang, S.E., Garcia-Molina, H.: Joint entity resolution. In: IEEE ICDE. Arlington, Virginia (2012)
277. Whang, S.E., Menestrina, D., Koutrika, G., Theobald, M., Garcia-Molina, H.: Entity resolution with iterative blocking. In: ACM SIGMOD, pp. 219–232. Providence, Rhode Island (2009)
278. Williams, G.J.: Rattle: a data mining GUI for R. The R Journal **1**(2), 45–55 (2009)
279. Winkler, W.: String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. In: Proceedings of the Section on Survey Research Methods, pp. 354–359. American Statistical Association (1990)
280. Winkler, W.E.: Using the EM algorithm for weight computation in the Fellegi-Sunter model of record linkage. Tech. Rep. RR2000/05, US Bureau of the Census, Washington, DC (2000)
281. Winkler, W.E.: Methods for record linkage and Bayesian networks. Tech. Rep. RR2002/05, US Bureau of the Census, Washington, DC (2001)
282. Winkler, W.E.: Record linkage software and methods for merging administrative lists. Tech. Rep. RR2001/03, US Bureau of the Census, Washington, DC (2001)
283. Winkler, W.E.: Approximate string comparator search strategies for very large administrative lists. Tech. Rep. RR2005/02, US Bureau of the Census, Washington, DC (2005)
284. Winkler, W.E.: Overview of record linkage and current research directions. Tech. Rep. RR2006/02, US Bureau of the Census, Washington, DC (2006)
285. Winkler, W.E.: Automatic estimation of record linkage false match rates. Tech. Rep. RR2007/05, US Bureau of the Census, Washington, DC (2007)
286. Winkler, W.E., Thibaudeau, Y.: An application of the Fellegi-Sunter model of record linkage to the 1990 U.S. decennial census. Tech. Rep. RR1991/09, US Bureau of the Census, Washington, DC (1991)
287. Winkler, W.E., Yancey, W.E., Porter, E.H.: Fast record linkage of very large files in support of decennial and administrative records projects. In: Proceedings of the Section on Survey Research Methods, pp. 2120–2130. American Statistical Association (2010)
288. Witten, I.H., Moffat, A., Bell, T.C.: Managing Gigabytes, 2 edn. Morgan Kaufmann (1999)
289. Xiao, C., Wang, W., Lin, X.: Ed-join: an efficient algorithm for similarity joins with edit distance constraints. Proceedings of the VLDB Endowment **1**(1), 933–944 (2008)
290. Yakout, M., Atallah, M., Elmagarmid, A.: Efficient private record linkage. In: IEEE ICDE, pp. 1283–1286 (2009)
291. Yakout, M., Elmagarmid, A., Elmeleegy, H., Ouzzani, M., Qi, A.: Behavior based record linkage. Proceedings of the VLDB Endowment **3**(1-2), 439–448 (2010)
292. Yan, S., Lee, D., Kan, M.Y., Giles, L.C.: Adaptive sorted neighborhood methods for efficient record linkage. In: ACM/IEEE-CS joint conference on Digital Libraries, pp. 185–194 (2007)
293. Yancey, W.E.: An adaptive string comparator for record linkage. Tech. Rep. RR2004/02, US Bureau of the Census (2004)
294. Yancey, W.E.: Evaluating string comparator performance for record linkage. Tech. Rep. RR2005/05, US Bureau of the Census (2005)
295. Yancey, W.E.: BigMatch: A program for extracting probable matches from a large file for record linkage. Tech. Rep. RRC2007/01, US Bureau of the Census (2007)
296. Yu, P., Han, J., Faloutsos, C.: Link Mining: Models, Algorithms, and Applications. Springer (2010)
297. Zaki, M., Ho, C.: Large-scale parallel data mining. Springer LNCS 1759 (2000)
298. Zhang, Y., Lin, X., Zhang, W., Wang, J., Lin, Q.: Effectively indexing the uncertain space. IEEE Transactions on Knowledge and Data Engineering **22**(9), 1247–1261 (2010)

299. Zhao, H.: Semantic matching across heterogeneous data sources. Communications of the ACM **50**(1), 45–50 (2007)
300. Zhu, J.J., Ungar, L.H.: String edit analysis for merging databases. In: KDD workshop on text mining, held at ACM SIGKDD. Boston (2000)
301. Zingmond, D., Ye, Z., Ettner, S., H., L.: Linking hospital discharge and death records – accuracy and sources of bias. Journal of Clinical Epidemiology **57**, 21–29 (2004)
302. Zobel, J., Dart, P.: Phonetic string matching: Lessons from information retrieval. In: ACM SIGIR, pp. 166–172. Zürich, Switzerland (1996)
303. Zobel, J., Moffat, A.: Inverted files for text search engines. ACM Computing Surveys **38**(2), 6 (2006)