German RLC

Record Linkage Bibliography

Basic Literature

Version 3.0

October 11, 2012

Tobias Bachteler, German Record Linkage Center

1. Overviews

Christen, P. 2012. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Heidelberg: Springer.

Elmagarmid, A., Ipeirotis, P. G., and Verykios, V. 2007. Duplicate record detection: a survey. *IEEE Transactions on Knowledge and Data Engineering* 19(1) 1–16.

Gill, L. E. 2001. *Methods for Automatic Record Matching and Linkage and Their Use in National Statistics*. Norwich: Office of National Statistics.

Herzog, T. N., Scheuren, F. J., and Winkler, W. E. 2007. *Data Quality and Record Linkage Techniques*. New York: Springer.

Herzog, T. N., Scheuren, F. J., and Winkler, W. E. 2010. Record linkage. *Wiley Interdisciplinary Reviews: Computational Statistics* 2(5) 535–543.

Newcombe, H. B. 1988. *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business*. Oxford: Oxford University Press.

Winkler, W. E. 1995. Matching and record linkage. B. G. Cox et al. (ed.) *Business Survey Methods*. New York: Wiley, pp. 355–384.

Winkler, W. E. 2004. Methods for evaluating and creating data quality. *Information Systems* 29(7) 531–550.

Winkler, W. E. 2009. Record linkage. D. Pfeffermann and C.R. Rao (ed.) *Handbook of Statistics 29A, Sample Surveys: Design, Methods and Applications*. Amsterdam: Elsevier, pp. 351–380.

2. Articles in Reference Books

Arasu, A. and Domingo-Ferrer, J. 2009. Record matching. L. Liu and M. T. Özsu (ed.) *Encyclopedia of Database Systems*. New York: Springer.

Domingo-Ferrer, J. 2009. Record linkage. L. Liu and M. T. Özsu (ed.) *Encyclopedia of Database Systems*. New York: Springer.

Fair, M. E. 2002. Record linkage. L. Breslow (ed.) *Encyclopedia of Public Health*. New York: Macmillan Reference/Gale Group.

Judson, D. H. 2004. Computerized record linkage and statistical matching. K. Kempf-Leonard (ed.) *Encyclopedia of Social Measurement*. Amsterdam: Elsevier.

3. Introductions

Clark, D. E. 2004. Practical introduction to record linkage for injury research. *Injury Prevention* 10(3) 186-191.

Hassard, T. H. 1986. Writing the book of life: medical record linkage. R. J. Brook et al. (ed.) *The Fascination of Statistics*. New York: Marcel Dekker, pp. 25-46.

Howe, G. R. 1998. Use of computerized record linkage in cohort studies. *Epidemiologic Reviews* 20(1) 112–121.

Smith, M. E. 1984. Record linkage: present status and methodology. *Journal of Clinical Computing* 13(2–3) 52–69.

4. Bibliographies and Literature Surveys

Elmagarmid, A., Ipeirotis, P. G., and Verykios, V. 2007. Duplicate record detection: a survey. *IEEE Transactions on Knowledge and Data Engineering* 19(1) 1–16.

Köpcke, H. and Rahm, E. 2010. Frameworks for entity matching: a comparison. *Data & Knowledge Engineering* 69(2) 197–210.

Machado, C. J. 2004. A literature review of record linkage procedures focusing on infant health outcomes. *Cadernos de Saúde Pública* 20(2) 362–371.

Winkler, W. E. 2011. Record linkage references (2011Jan15).

5. Preprocessing

Christen, P. 2012. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Heidelberg: Springer, Chapter 3: Data Pre-Processing.

Herzog, T. N., Scheuren, F. J., and Winkler, W. E. 2007. *Data Quality and Record Linkage Techniques*. New York: Springer, Chapter 10: Standardization and Parsing.

Low, W. L., Lee, M. L., and Ling, T. W. 2001. A knowledge-based approach for duplicate elimination in data cleaning. *Information Systems* 26(8) 585–606.

Smith, M. E. 1985. Record-keeping and data preparation practices to facilitate record linkage. B. Kilss and W. Alvey (ed.) *Record Linkage Techniques 1985. Proceedings of the Workshop on Exact Matching Methodologies: 9–10 May 1985; Arlington, VA*. Washington, DC: Department of the Treasury, Internal Revenue Service, Statistics of Income Division, pp. 321–326.

Winkler, W. E. 1985. Preprocessing of lists and string comparison. B. Kilss and W. Alvey (ed.) *Record Linkage Techniques 1985. Proceedings of the Workshop on Exact Matching Methodologies: 9–10 May 1985; Arlington, VA*. Washington, DC: Department of the Treasury, Internal Revenue Service, Statistics of Income Division, pp. 181–187.

6. Blocking Methods

Baxter, R., Christen, P., and Churches, T. 2003. A comparison of fast blocking methods for record linkage. CIMS Technical Report 03/139, CSIRO Mathematics, Informatics and Statistics.

Christen, P. 2012. A survey of indexing techniques for scalable record linkage and deduplication. *IEEE Transactions on Knowledge and Data Engineering* 24(9) 1537–1555.

Christen, P. 2012. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Heidelberg: Springer, Chapter 4: Indexing.

Herzog, T. N., Scheuren, F. J., and Winkler, W. E. 2007. *Data Quality and Record Linkage Techniques*. New York: Springer, Chapter 12: Blocking.

7. Rule-Based Record Linkage

Hernández, M. A. and Stolfo, S. S. 1998. Real-world data is dirty: data cleansing and the merge/purge problem. *Data Mining and Knowledge Discovery* 2(1) 9–37.

Low, W. L., Lee, M. L., and Ling, T. W. 2001. A knowledge-based approach for duplicate elimination in data cleaning. *Information Systems* 26(8) 585–606.

Whang, S. and Garcia-Molina, H. 2010. Entity resolution with evolving rules. *Proceedings of the VLDB Endowment* 3(1) 1326–1337.

8. Distance-Based Record Linkage

Arasu, A., Chaudhuri, S., and Kaushik, R. 2008. Transformation-based framework for record matching. *Proceedings of the 24th International Conference on Data Engineering: 7–12 April 2008; Cancún, Mexico*. Los Alamitos, CA: IEEE, pp. 40–49.

Christen, P. 2012. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Heidelberg: Springer, Chapter 5: Field and Record Comparison.

Monge, A. E. and Elkan, C. P. 1996. The field-matching problem: algorithms and applications. E. Simoudis, J. Han, and U. Fayyad (ed.) *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining: 2–4 August 1996; Portland*. Menlo Park, CA: AAAI Press, pp. 267–270.

Whang, S. and Garcia-Molina, H. 2010. Entity resolution with evolving rules. *Proceedings of the VLDB Endowment* 3(1) 1326–1337.

8.1 Overviews of String Similarity Functions

Bilenko, M. et al. 2003. Adaptive name matching in information integration. *IEEE Intelligent Systems* 18(5) 16–23.

Christen, P. 2006. A comparison of personal name matching: techniques and practical issues. S. Tsumoto et al. (ed.) *Proceedings of the 6th IEEE International Conference on Data Mining - Workshops: 18 December 2006; Hong Kong.* Los Alamitos, CA: IEEE, pp. 290–294.

Hall, P. A. V. and Dowling, G. R. 1980. Approximate string matching. *ACM Computing Surveys* 12(4) 381–402.

Stephen, G. A. 1994. *String Searching Algorithms.* Singapore: World Scientific.

8.2 Comparisons of String Similarity Functions

Bilenko, M. et al. 2003. Adaptive name matching in information integration. *IEEE Intelligent Systems* 18(5) 16–23.

Christen, P. 2006. A comparison of personal name matching: techniques and practical issues. S. Tsumoto et al. (ed.) *Proceedings of the 6th IEEE International Conference on Data Mining - Workshops: 18–22 December 2006; Hong Kong.* Los Alamitos, CA: IEEE, pp. 290–294.

Yancey, W. E. 2005. Evaluating string comparator performance for record linkage. Technical Report RSS2005/05. Statistical Research Division, U.S. Census Bureau, Washington, DC.

8.3 Numerical Similarity Functions

Agrawal, R. and Srikant, R. 2002. Searching with numbers. D. Lassner, D. De Roure, and A. Iyengar (ed.) *Proceedings of the 11th International World Wide Web Conference: 7–11 May 2002; Honolulu*. New York: ACM, pp. 420–431.

9. Probabilistic Record Linkage

Fellegi, I. P. and Sunter, A. B. 1969. A theory for record linkage. *Journal of the American Statistical Association* 64(328) 1183–1210.

Herzog, T. N., Scheuren, F. J., and Winkler, W. E. 2007. *Data Quality and Record Linkage Techniques*. New York: Springer, Chapter 8: Record Linkage – Methodology.

Jaro, M. A. 1989. Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association* 84(406) 414–420.

9.1 Adjustment of Weights for String Similarities

Herzog, T. N., Scheuren, F. J., and Winkler, W. E. 2007. *Data Quality and Record Linkage Techniques*. New York: Springer, Chapter 13: String Comparator Metrics for Typographical Error.

Winkler, W. E. 1990. String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. *Proceedings of the Section on Survey Research Methods.* American Statistical Association, pp. 354–359.

Yancey, W. E. 2005. Evaluating string comparator performance for record linkage. Technical Report RSS2005/05. Statistical Research Division, U.S. Census Bureau, Washington, DC.

9.2 Frequency Weights

Newcombe, H. B. 1988. *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business*. Oxford: Oxford University Press, Chapter 9: Converting 'global' odds to 'specific' odds.

Winkler, W. E. 1989. Frequency-based matching in the Felligi-Sunter model of record linkage. *Proceedings of the Section on Survey Research Methods.* American Statistical Association, pp. 778–783.

Yancey, W. E. 2000. Frequency-dependent probability measures for record linkage. *Proceedings of the Section on Survey Research Methods.* American Statistical Association, S. 752–757.

9.3 Parameter Estimation

Herzog, T. N., Scheuren, F. J., and Winkler, W. E. 2007. Data Quality and Record Linkage Techniques. New York: Springer, Chapter 9: Estimating the Parameters of the Fellegi–Sunter Record Linkage Model.

Winkler, W. E. 1993. Improved decision rules in the Fellegi-Sunter model of record linkage. *Proceedings of the Section on Survey Research Methods.* American Statistical Association, pp. 274–279.

Winkler, W. E. 1988. Using the EM algorithm for weight computation in the Fellegi-Sunter model of record linkage. *Proceedings of the Section on Survey Research Methods.* American Statistical Association, pp. 667–671.

9.4 Threshold Selection/Error Rate Estimation

Belin, T. R. and Rubin, D. B. 1995. A method for calibrating false-match rates in record linkage. *Journal of the American Statistical Association* 90(430) 694–707.

Bishop, G. and Khoo, J. 2006. Methodology of evaluating the quality of probabilistic linking. *Proceedings of Statistics Canada Symposium. Methodological Issues in Measuring Population Health: 1–3 November 2006; Gatineau, Canada*. Ottawa: Statistics Canada.

Winkler, W. E. 1995. Matching and record linkage. B. G. Cox et al. (ed.) *Business Survey Methods*. New York: Wiley, pp. 355–384, Section 20.6: Estimation of Error Rates and Adjustment for Matching Error.

Winkler, W. E. 2006. Automatically estimating record linkage false match rates. *Proceedings of the Survey Research Methods Section.* American Statistical Association, pp. 3863–3870.

10. 1-1 Assignment

Jaro, M. A. 1989. Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association* 84(406) 414–420.

Winkler, W. E. 1994. Advanced methods for record linkage. Technical Report RR94/05. Statistical Research Division, U.S. Census Bureau, Washington, DC.

11. Bias due to Imperfect Record Linkage

Chesher, A. and Nesheim, L. 2006. Review of the literature on the statistical properties of linked datasets. Technical Report 3. Department of Trade and Industry, London.

Lahiri, P. and Larsen, M. D. 2005. Regression analysis with linked data. *Journal of the American Statistical Association* 100(469) 222–230.

Ridder, G. and Moffitt, R. 2007. The econometrics of data combination. J. J. Heckman and E. E. Leamer (ed.) *Handbook of Econometrics.* Amsterdam: Elsevier, pp. 5469–5547.

12. Record Linkage Software

ESSnet Statistical Methodology Project on Integration of Survey & Administrative Data 2011. Deliverable WP3: software tools for integration methodologies.

Herzog, T. N., Scheuren, F. J., and Winkler, W. E. 2007. *Data Quality and Record Linkage Techniques*. New York: Springer, Chapter 19: Review of Record Linkage Software.

13. Privacy-Preserving Record Linkage

Christen, P. 2012. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Heidelberg: Springer, Chapter 8: Privacy Aspects of Data Matching.

Hall, R. and Fienberg, S. E. 2010. Privacy-preserving record linkage. J. Domingo-Ferrer and E. Magkos (ed.) *Proceedings of the 2010 International Conference on Privacy in Statistical Databases: 22–24 September 2010; Corfu, Greece*. Berlin: Springer, Berlin, pp. 269–283.

Karakasidis, A. and Verykios, V. S. 2011. Advances in Privacy Preserving Record Linkage. T. Matsuo and T. Fujimoto (ed.) *E-Activity and Intelligent Web Construction: Effects of Social Design*. Hershey, PA: Information Science Reference, Hershey, pp. 22–34.