Record Linkage for Genealogical Databases

Dallan Quass
Department of Computer Science
Brigham Young University
Provo, Utah 84602 USA
1-801-240-6930
quass@byu.edu

Paul Starkey
Family and Church History Department
Church of Jesus Christ of Latter-day Saints
Salt Lake City, Utah 84150 USA
1-801-240-4451
starkeypd@ldschurch.org

ABSTRACT
In this paper we describe past experience and outline current directions in performing record linkage over large genealogical databases.

1. INTRODUCTION AND MOTIVATION
Record linkage is the problem of identifying multiple records that refer to the same real-world entity. In genealogical databases, it is the problem of identifying when individuals situated in different pedigrees refer to the same real-world individual. Being able to link records in genealogical databases has value to people engaged in genealogical research because it condenses search results and helps people identify when their work overlaps with the research of others. It also has value to medical researchers trying to understand the hereditary nature of cancer, heart disease, and other illnesses.

Unlike most record linkage problems, record linkage in genealogical databases usually allows one to utilize a broad range of features, since records are often situated in the context of pedigrees. For many individuals within a pedigree, dates and locations of birth, marriage, and death are usually available, as well as information about children, spouses, siblings, and parents. Often individuals within a pedigree are identified through different vital records by different genealogical researchers. Furthermore, the linkage problem can be cast as a graph-matching problem, since the decision to link (or not to link) two individuals influences the decision to link individuals related to them [5]. Finally, any record linkage problem has the issue of determining the similarity of different names [3,6], determining for example the probability that Peg, Peggy, and Margaret name the same person. Linked genealogical databases can provide insight into name similarity since we can identify the various names associated with linked records. Since the linking of genealogical databases should be quite accurate due to the broad range of features and the ability to use graph-matching concepts, one could be reasonably confident in the names that are found to be similar.

ACM SIGKDD '03 Workshop on Data Cleaning, Record Linkage, and Object Consolidation, August 24-27, 2003, Washington, DC.

2. PRIOR EXPERIENCE LINKING GENEALOGICAL DATA
The Church of Jesus Christ of Latter-day Saints has been involved in genealogical research for over a century. The doctrinal foundations of the Church teach that families do not end at death, but instead are the basis of society in the world beyond death. With that understanding, members of the Church regard it as a privilege and obligation to seek out their forbearers.

The Church began gathering genealogical records in 1894 and now maintains the world's largest repository of genealogical resources. These resources include:

- *International Genealogical Index* (IGI) – 400 million records of deceased individuals. Many records contain information about the parents or spouse of the individual as well. Forty-seven percent come from original vital records (e.g., births, marriages, deaths); the remaining records come from compiled sources such as pedigrees. As pedigrees were placed into the database, only immediate-family husband-wife and parent-child relationships were preserved; extended family relationships across multiple generations were not maintained in the database.
- *Ancestral File* (AF) – 256,000 pedigrees submitted between 1984 and 1999. Submissions were kept as pedigree data structures so that both immediate and extended family relationships were maintained. An attempt was made to automatically link and merge records from different pedigrees, which resulted in a database of 40 million individuals.
- *Pedigree Resource File* (PRF) – 60 million individuals in approximately 41,000 pedigrees submitted after June 1999. No attempt has been made to link records.
- *Granite Mountain Record Vault* (GMRV) – over two billion pages of historical documents related to genealogy (e.g., parish, census, and vital records) stored on 2.2 million rolls of microfilm.

The IGI, AF, and PRF databases are searchable online at http://www.familysearch.org, and films in the GMRV collection are orderable through a network of local family history centers.

Record linkage was performed on pedigree submissions to AF in an attempt to show people where their pedigrees overlapped, giving them an opportunity to collaborate with other submitters on future research. Individuals in different pedigree submissions that were linked (identified as duplicates) by the record linkage algorithm were merged into a single individual in the AF database, with the most-recently submitted data taking precedence in the merge.

The initial record linkage algorithms were developed heuristically based upon matching Soundex codes on names and exact matching on date and place fields for some event (birth, marriage, or death). Once a pair of duplicates was linked, additional, less-stringent rules were used to link related family members.

A few years later, statistical record linkage based upon a Fellegi-Sunter algorithm [2,4] was added to the process. The features used in the statistical algorithm were based upon name and event information. Features based upon family relationships such as names of parents or spouse were not used. Independent feature weights and thresholds were derived for various world regions. For example, individuals born in Scandinavian countries were linked using one set of weights and thresholds; individuals born in the USA were linked using another set of weights and thresholds. Individuals whose birthplace was unknown were linked using a "generic world" set of weights and thresholds.

The implementation of statistical record linkage was generally effective at matching individuals. The heuristic family relationship linking rules were again used in a post-process to link other individuals in the family who were "poorly identified" – those who did not contain sufficient identifying name and event information to be linked using the statistical algorithm. Additional rule-based filtering was also added to this hybrid to prevent false links, such as those resulting from same-named siblings born within one or two years. (It was a common practice a century ago to reuse given names of children who died in infancy by giving them to the next child born if he/she had the same gender.)

In retrospect, combining the early heuristic rules with the statistical algorithm was a mistake. In later record linkage projects the statistical algorithm was found to perform better without the heuristic family-member matching and post-filtering rules. Because of issues surrounding incorrect linking in AF, a new PRF database was established and new submissions were directed to that database.

3. CURRENT DIRECTION – CREATING LABELED DATASETS

The Church recently embarked upon an effort to combine the IGI, AF, and PRF databases into a single database. We plan to link records initially in batch mode as we populate the database from these sources, and then interactively as new pedigrees are submitted. Linking records within the IGI is similar to a traditional record-linkage problem, since little relationship information is included. A difference is that we hope to reconstruct multi-generation pedigree structures from the data in IGI that was originally submitted in pedigree format. Linking records within AF and PRF and between these two databases and the IGI records that have been placed into pedigrees allows for the broad range of features and graph-matching opportunities discussed in Section 1.

As a first step in linking the records, we plan to create sample datasets from each of the three sources and to manually label linked records within and between the samples. We again intend to use a Fellegi-Sunter-based statistical linking algorithm. We hope to augment the algorithm with name-similarity and location-proximity metrics. In addition, we plan to use features based upon family relationships and hope to integrate graph-based linking techniques into the statistical algorithm.

A basic statistical record-linkage algorithm will be run over the datasets. Standard blocking techniques will be used to reduce the number of pairs of individuals to consider. Once the algorithm has been run, random sampling will be used to determine the lower-confidence threshold, below which all pairs will be set automatically to "not linked." Likewise, sampling will be used to determine the upper-confidence threshold, above which all pairs will be set automatically to "linked." Pairs whose linking score falls between the lower and upper thresholds will be reviewed by a team of expert genealogists to determine which are true links. By manually reviewing all pairs that fall between the upper and lower thresholds, we expect the recall and precision of the labeled datasets to be high. As we identify additional links over time, we will add those links to the labeled datasets. Once the labeled datasets have been created, we will use them in the development and evaluation of more advanced algorithms.

We are considering making available our labeled sample PRF dataset as a benefit to others developing and evaluating record linkage algorithms. We expect that this dataset would contain approximately 600 pedigrees chosen at random totaling nearly one million individuals, with the links labeled. Depending upon the level of interest, other datasets could follow. Anyone interested in the possibility of obtaining labeled datasets for research purposes should contact Dallan Quass.

4. FUTURE WORK

In addition to identifying potential duplicates in pedigrees, another area for record linkage research involves linking individuals found in vital, census, and other original source records for a geographic region and time period into probable pedigree structures. This type of work is termed *family reconstruction*. Family reconstruction is possible when (a) several types of records (e.g., birth, marriage, and death certificates) are available for a particular geographic region and time period, (b) the records overlap in the information they contain (e.g., birth certificates list names of both parents and marriage certificates list parents of both spouses), (c) most of the records are available in electronic form, and (d) it is known that people did not often emigrate from or immigrate into the region during the time period.

Family reconstruction could provide a possible starting point for beginning genealogists and medical researchers. Prior work has been done in this area ([1] for example), but there is room for much improvement. We intend to address this problem as future work.

5. ACKNOWLEDGEMENTS

The authors wish to thank David Barss, Tom Creighton, Chris Cummings, Kevin Johnson, Ryan Knight, Heath Nielson, Randy Wilson, and especially Nancy NeSmith for their comments, suggestions, and continuing work on the problem.

6. REFERENCES

[1] Bloothooft, G. Multi-Source Family Reconstruction. History and Computing, 7,2 (1995), 90-103.

[2] Fellegi, I.P., and Sunter, A.B. A Theory for Record Linkage. Journal of the American Statistical Association, 64 (1969), 1183-1210.

[3] Jaro, M.A. Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. Journal of the American Statistical Association, 89 (1989), 414-420.

[4] Newcombe, H.B., Kennedy, J.M., Axford, S.J., and James, A.P. Automatic Linkage of Vital Records. Science, 130 (1959), 954-959.

[5] Wilson, R. Graph-Based Remerging of Genealogical Databases. In Proceedings of the 2001 Family History Technology Workshop (Provo UT, 2001), 4-6. http://www.fht.byu.edu/workshop01/final/Wilson.pdf

[6] Winkler, W.E. Advanced Methods of Record Linkage. American Statistical Association, Proceedings of the Section of Survey Research Methods, (1994), 467-472.