A Metric-Based Machine Learning Approach to Genealogical Record Linkage

S. Ivie, G. Henry, H. Gatrell and C. Giraud-Carrier
Department of Computer Science, Brigham Young University

Abstract

Genealogical Record Linkage (GRL) is the process of determining whether two pedigrees refer to the same base individual. Unlike other record linkage problems, GRL datasets have a large number of attributes that frequently are sparsely populated with no definitive limit. A metric-based, machine learning approach has been developed. In this approach, innovative comparison metrics were developed for the three basic types of data: names, dates and locations. In addition, two more advanced comparisons were developed to handle one-to-many relationships (e.g., an individual may have 0 to an unknown number of children). Using these metrics and Clementine's C5.0 decision tree learning algorithm (with costs and boosting), high levels of accuracy, precision, and recall were achieved on a large post-blocking, standardized database.

Keywords: *record linkage, data linkage, duplicate record detection, de-duplication, data integration and matching, database merging, record linkage in sparsely populated databases, date comparison metric, location comparison metric, attribute grouping*

1 Introduction

Record linkage consists of discovering duplicate records within a data collection, or combining multiple overlapping data collections such that records that are believed to refer to the same entity are treated as a single entity. Record linkage has many applications, one of which is genealogical record linkage (GRL). In GRL, a record is a pedigree consisting of a base individual, his/her siblings, spouse, progeny and ancestry, all with basic information about major lifetime events including dates and places. GRL primarily focuses on determining whether or not two pedigrees refer to the same base individual. Each pedigree in such a comparison may be very unique due to spelling errors, data entry errors, variations between two or more databases, missing values, etc. As such, GRL considers more than exact-match pedigrees; it considers pedigrees that may differ drastically, but in actuality refer to the same individual.

GRL is significant to genealogical research because it consolidates and links numerous databases, resulting in condensed search results that have a broad range of highly related information. GRL also helps genealogical researchers identify where their work overlaps with the work of others. Furthermore, GRL has application in medical genetics where researchers identify the heredity of diseases such as cancer and heart conditions using medical pedigree charts [6]. GRL differs from other record linkage problems in the quantity and nature of the attributes used to represent entities. Where most record linkage projects have records that consist of a small and finite number of densely populated attributes, GRL tends to have a large number of attributes that are generally sparsely populated and may be multi-valued. For example, in a

pedigree an individual can have multiple spouses (due to remarriage, etc.), many children, many siblings, and a vast posterity and ancestry, each with numerous attributes.

This paper presents MBGRL, a metric-based machine learning approach to genealogical record linkage. A set of effective metrics is designed for each basic data type, as well as for multi-valued attributes. These are used in turn to train a decision tree learning algorithm for the task of record linkage. Results on a large genealogical database show high precision and recall.

2 Data Used

The genealogical database used in our experiments was provided by the Family and Church History Department (FCHD) of The Church of Jesus Christ of Latter-day Saints. The database consists of a set of pedigree comparisons, where each pedigree comparison is labeled as either being a "match" or "non-match." Blocking on this database was preformed previously by the FCHD so that only pairs that are very similar are left in the provided database. The distribution of matches to non-matches is approximately 1:3 (i.e., 1 match for every 3 non-matches), or approximately 25% matches. The database has also been heavily standardized, meaning it has been through many data cleaning and attribute-level reconciliation algorithms that have made every attribute conform to some standard form. For example, all abbreviations and misspellings in the city attribute have been converted to actual full, unabbreviated city names.

The database consists of names of people (e.g., "Jane Doe"), relationships (father, mother, sibling, child, spouse), and events (birth, christening, marriage, burial, etc.). An event consists of a date and a place. For evaluation purposes, the database was split into two sub-sets: a training set consisting of two-thirds of the data, and a test set consisting of the remaining third. The test set was separated from all development, evaluation, and testing, and was only used to verify results at the end.

3 Developing and Choosing Comparison Metrics

Most genealogical record linkage problems involve comparisons among primarily four types of basic data types: name, gender, date, and location. An additional complex comparison is also needed to handle one-to-many relationships (e.g., an individual may have 0 to an unknown number of children). A wide variety of metrics were tested in each of these five comparison areas. To determine which metrics was most adequate on each data type, a metric performance evaluation was performed, as follows.

3.1 Metric Performance Evaluation Criteria

All metrics in a comparison category (name, date, location, etc.) were compared with each other using the following three criteria.

Information Gain. The formula for information gain is given in Equation 1. Information gain measures how well a given attribute (consisting of the results of a metric

comparison) separates the training data according to its target classification (match, mismatch) [4].

Let:

$$Entropy(S) \equiv -p_{\oplus} \log_2 p_{\oplus} - p_{\ominus} \log_2 p_{\ominus}$$

Where:

S is the collection of results a particular comparison metric generates with their associated target (match, mismatch), $p_{\oplus}$ is the proportion of matches, and $p_{\ominus}$ is the proportion of mismatches.

Then:

$$Gain(S,A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Where:

Values(A) is the set of all possible values for attribute *A*, and S_v is the subset of *S* where attribute *A* has value *v*, i.e., $S_v = \{s \in S | A(s) = v\}$.

Equation 1: Information Gain

F-score. The formula for F-score is given in Equation 2. The F-score tries to combine precision and recall into a single measure. F-scores were calculated using 10-fold cross-validation [4].

Let:

$$Precision = \frac{TP}{TP + FN} * 100\%$$

$$Recall = \frac{TP}{TP + FP} * 100\%$$

Where:

TP is the number of true positives (number of correctly labeled matches), *FN* is the number of false negatives (number of matches incorrectly labeled as mismatches), and *FP* is the number of false positives (number of mismatches incorrectly labeled as matches).

Then:

$$Fmeasure = 2 \times \frac{\text{recall} \times \text{precision}}{recall + precision}$$

Equation 2: F-measure

Overall Accuracy. The overall accuracy is simply the ratio of the number of correctly labeled pairs to the total number of pairs in the training set. Overall accuracy was computed using 10-fold cross-validation.

A metric was considered to be superior only if it outperformed the other metrics in its category on all 3 of the above criteria

3.2 Metric Performance Evaluation Results

A number of metrics were tested for each data type and evaluated based on the criteria of section 3.1. The following comparison metrics were found to be superior in their respective comparison groups:

Name Comparison Metric. During the blocking stage, name comparisons were optimized for speed, while maintaining accuracy. Once the data was past the blocking, a more time intensive name comparison metric was used. A weighted ensemble of string metrics — Monge-Elkan [5], Jaro-Winkler [3] and Soundex [8] — resulted in the highest performance evaluation improvement. These comparators have consistently shown high accuracy on name matching tasks [2,7] and when combined into a weighted ensemble show even higher accuracy for our dataset.

Date Comparisons Metric. When two dates (in the same year and month) are compared, a similarity score is calculated primarily according to the absolute value of the difference in number of days between the two dates. For example 1 June 1800 and 10 June 1800 would conceptually result in a score of 9 because there is a 9-day difference between the two dates. A score of 0 means an exact match and a high score implies a low probability that the two dates match.

Allowances are also made to the score according to common data recording errors. For example, the comparison of 21 June 1800 and 12 June 1800 will score slightly lower (having a greater probability) than 10 June 1800 and 19 June 1800, because the common error of reversing date digits implies a slightly higher probability of being a match (i.e., it is more likely that “21” matches “12” than that “10” matches “19”, even though the difference in number of days is the same).

Location Comparison Metric. As stated earlier, the locations in the dataset have already been standardized. This means that misspellings, variations and abbreviations have been previously resolved to actual locations (past and present). Location strings are compared initially to see if there is an exact match, assuming all four parts of a location are present (i.e., city, county, state, and country). If they are not a match, traditional string comparison metrics are rendered useless because the location names have already been standardized. For example, it makes no sense to compare the string similarity of “Manhattan” and “New York City.” Instead, a physical distance metric was created.

Using Yahoo Maps online services, literal distances are calculated between two locations (cities). Using a physical distance metric allows for greater sensitivity of determining

common data entry errors. For example, one "birthplace" may erroneously list a larger city like Salt Lake, rather than the actual suburb, like Sandy. Another common location discrepancy exists between a pedigree that lists the city of the hospital an individual was born in, and another pedigree lists the city the individual's parents lived in when the individual was born (e.g., someone was born at the Provo hospital, but lives in and is from Orem).

Over a period of time, a database was created with every unique location in the database and its corresponding geo-coordinates. Distances were calculated using Equation 3. Over 85% of the locations could be resolved to coordinates (the remainder are cities that no longer exist, are not yet indexed by yahoo, etc).

Let:

$$D = r \times [\sin La_1 \times \sin La_2 + \cos La_1 \times \cos La_2 \times \cos(Lo_2 - Lo_1)]$$

Where:
D is the distance in kilometers, *r* is the radius of the earth in kilometers, *La* is the latitude in radians, and *Lo* is the longitude in radians.

Equation 3: Distance between geo-coordinates

This standardized, literal-distance location metric shows minor improvements in performance. This metric is also "future-friendly," as many genealogy programs allow users to enter GPS coordinates for locations, such as grave sites [1].

One-To-Many Comparisons. In GRL, there are many comparisons that have one-to-many relationships. For example, a person may remarry multiple times and thus have a number of spouses; a person may also have a large number of children, many siblings, etc. The approach that proved best for children, siblings and spouses is a form of "winner take all" method, as follows.

Each child (respectively, sibling or spouse) in one pedigree is compared to each child (respectively, sibling or spouse) in the other pedigree. The name of the individual (child, sibling or spouse), the name of the individual's spouse, and their respective events are all weighted and combined into a single standardized score. The pair-wise comparison that results in the highest score is the score that is used for the comparison of all the children (respectively, siblings or spouses).

4 Experiments and Results

For each set of pedigrees in the database, a record is generated and output using the metric specific to each attribute. SPSS Clementine's C5.0 decision tree learning algorithm is then run on this output, with disproportional weighting towards false negatives (type II error), due to the bias

inherent in the data (the 1:3 ratio of non-matches to matches). Boosting is also used to improve accuracy, and aggressive pruning is applied to preserve generality.

As mentioned in section 2, our data is split into two subsets, one for training and one for testing. After the best comparison algorithms were found using ten-fold cross-validation on the training data, a C5.0 decision tree model was induced from it. The test set was then "blindly" run through the resulting decision tree (the first time that set was used for any purpose). The combination of our metric-based algorithms resulted in high accuracy, F-score, precision, and recall as shown below. The upper table is the confusion matrix (0 stands for Mismatch and 1 stands for Match), whilst the lower table summarizes the main quantities of interest.

		PREDICTED	
		0	**1**
ACTUAL	**0**	3719	65
	1	102	1276

SUMMARY

Actual Matches:	1,378
Actual Mismatches:	3,784
Total Comparisons:	5,162
Match/Mismatch Ratio:	0.3642
Accuracy:	0.9676
Precision:	0.9260
Recall:	0.9515
F-score:	0.9386

5 Comparison to Other Methods Using the Same Data

At least two other groups of people have performed genealogical record linkage on the dataset used here. The first used MAL 4:6, a structured neural network created previously by our colleagues at the Brigham Young University Data Mining Lab [9,10]. The second comparison was performed by colleagues from the FCHD, using a combination of hand-crafted rules and machine learning techniques.

Figures 2 and 3 show precision-recall charts for all three approaches. In Figure 2, all available attributes are being used in the comparison, whilst in Figure 3, only the attributes of the base individuals are considered. MBGRL performs very well overall. The precision-recall curve is only slightly below that of the FCHD in Figure 2, which is rather promising as MBGRL is fully automated and does not rely on any human-generated rules.

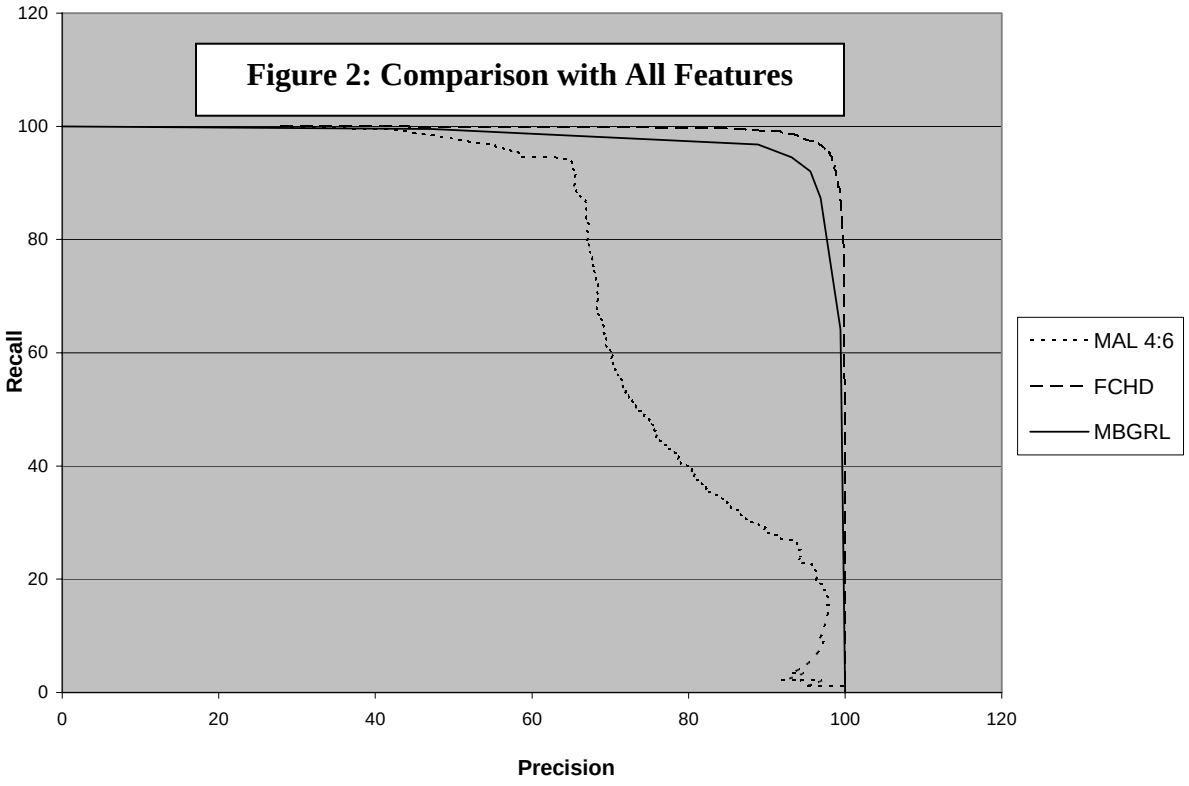
Figure 2: Comparison with All Features
120
100
80
60
40
20
0
Recall
0
20
40
60
80
100
120
Precision
MAL 4:6
FCHD
MBGRL

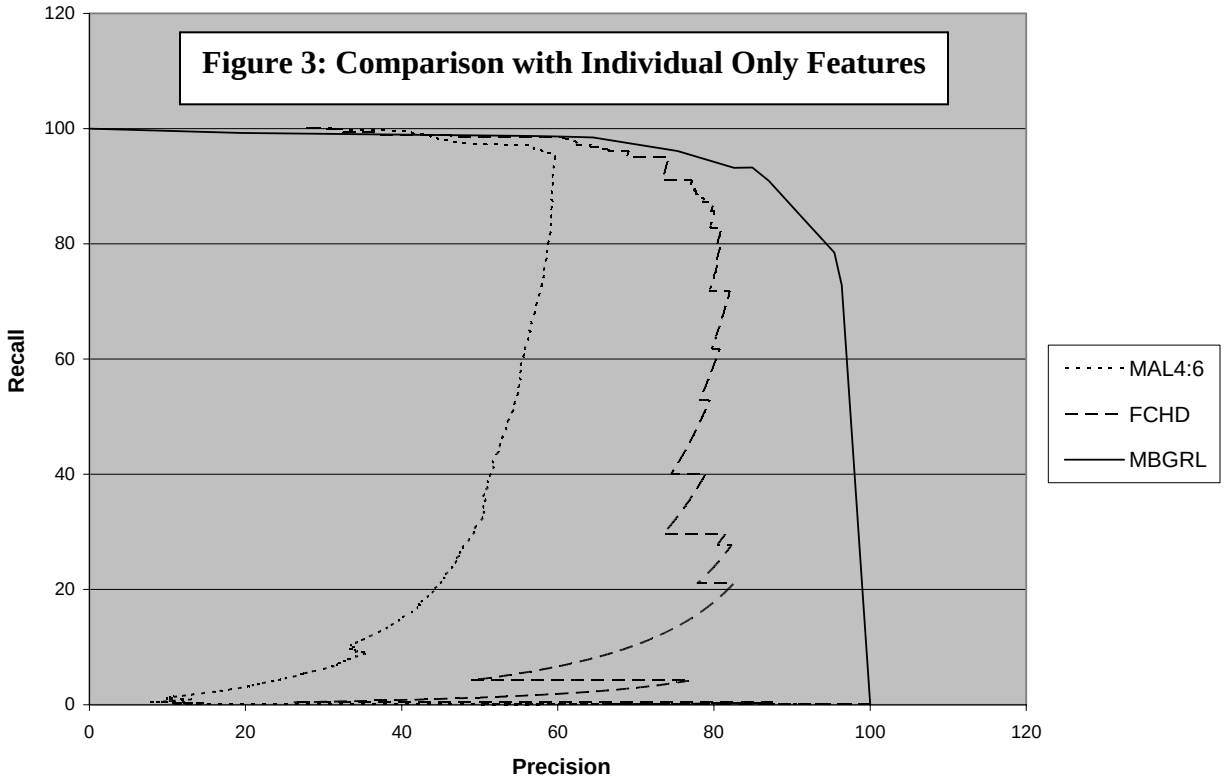
Figure 3: Comparison with Individual Only Features
120
100
80
60
40
20
0
Recall
0
20
40
60
80
100
120
Precision
MAL4:6
FCHD
MBGRL

6 Conclusion

A metric based machine learning approach to genealogical record linkage has been presented. By evaluating various metrics for each of the comparison types of genealogical data (name, gender, location, date) and multi-valued attributes, high-performing comparison metrics well suited for our genealogical data were selected. When these metrics were combined into a SPSS Clementine C5.0 decision tree learning algorithm, high levels of accuracy, precision, and recall were achieved. These results are encouraging. They exceed those obtained by previous automated approaches, and are comparable to approaches optimized with hand-crafted rules.

Acknowledgments

Data for our experiments was graciously provided by the Family & Church History Department of the Church of Jesus Christ of Latter-day Saints. We express special thanks to Jun won Lee for assistance with SPSS Clementine, as well as other members of the BYU Data Mining Lab for insightful discussions on record linkage.

References

[1] Booth, M. T. (2006). Enhancing Your Genealogy Using GPS *(Slideshow)*. Available online from http://www.personalhistorian.com/Support/EnhancingYourGenealogyWithGPS.pdf. Retrieved March 1, 2007.

[2] Cohen, W.W., Ravikumar, P. and Fiendberg, S.E. (2003). A Comparison of String Distance Metrics for Name-Matching Tasks. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*.

[3] Jaro, M.A. (1995). Probabilistic Linkage of Large Public Health Data File. *Statistics in Medicine*, **14**:491-498.

[4] Mitchell, T.M. (1997). *Machine Learning.* New York: McGraw-Hill, pp. 55-58.

[5] Monge, A. and Elkan, C. (1996). The Field-Matching Problem: Algorithm and Applications. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, 267-270.

[6] NeSmith, N.P. (1999). *Record Linkage Techniques -- 1997: Proceedings of an International Workshop and Exposition*, March 20-21, 1997, Arlington, VA. Washington, DC: Federal Committee on Statistical Methodology Office of Management and Budget, p. 358.

[7] Pfeifer, U., Poersch, T. and Fuhr, N. (1996). Retrieval Effectiveness of Proper Name Search Methods. *Information Processing and Management*, **32**(6):667-679.

[8] Phua, C., Lee, V. and Smith, K. (2006). The Personal Name Problem and a Recommended Data Mining Solution. *Encyclopedia of Data Warehousing and Mining (2nd Edition).*

[9] Pixton, B. and Giraud-Carrier, C. (2006). Using Structured Neural Networks for Record Linkage. In *Proceedings of the 6th Annual Workshop on Technology for Family History and Genealogical Research*.

[10] Pixton, B. and Giraud-Carrier, C. (2005). MAL4:6 - Using Data Mining for Record Linkage. In *Proceedings of the 5th Annual Workshop on Technology for Family History and Genealogical Research*.