MAL4:6 - Using Data Mining for Record Linkage

Burdette Pixton and Christophe Giraud-Carrier
Department of Computer Science, Brigham Young University
Provo, UT 84602

Abstract

This paper presents a first attempt at using pedigree-based data to improve record linkage. It describes a composite metric for similarity and a mechanism to extract relevant generational features. Results on a large data set demonstrate promise.

1 Introduction

Record linkage is the process of integrating information from two or more independent sources, ensuring that records believed to represent the same object are matched and treated as a single entity. In the context of genealogical databases, record linkage seeks specifically to identify whether or not individuals belonging to different pedigrees refer to the same person [11]. As the custodian of the largest source of genealogical data, the Church of Jesus Christ of Latter-day Saints' Family History Department (FHD) has a keen interest in accurate record linkage to (a) guarantee the integrity and overall quality of the available data, and (b) facilitate researchers' work by merging together complementary information.

The current record linkage system used by the FHD relies on a block-score mechanism based on sophisticated, hand-crafted comparison rules (e.g., if the birth year matches within some tolerance, a positive score is awarded; if it differs, a negative score). This approach, known as the Probabilistic Record Linkage formula was introduced by Newcomb [9] and formalized by Fellegi & Sunter [3].

Hand-crafted rules rely on increasingly specific and complex features generally designed through trial-and-error. Data mining techniques, which exploit pedigree charts more directly by considering both individual data as well as family relationships, offer a promising, more reliable alternative to discover relevant features, thereby improving record linkage efficiency and accuracy. This paper presents one such approach, MAL4:6 V0.1 (Mining And Linking for Successful Information eXchange).

The paper is organized as follows. Section 2 presents a brief study of various similarity metrics and proposes a composite metric for genealogical records. Section 3 shows how scorecards are used to extract relevant features from pedigree charts. Section 4 reports on the use of graph-based matching for record linkage. Finally, section 5 concludes the paper.

2 Metric Selection

Genealogical records contain data of heteregeneous types, such as gender, names, dates and locations, which potentially exhibit different behaviors under different similarity metrics. Yet, most record matching systems apply a single metric uniformly across all types.

In the spirit of [5], we seek to design a composite similarity metric that capitalizes on the inherent heterogeneity of genealogical records. We consider Soundex [15], Jaro's metric [6, 7], Jaro-Winkler's metric [13], Dice's approach [2, 12], binary discrimination (i.e., same vs different) and the inverse of the 1-norm (i.e., 1 over the absolute value of the difference; 1 if values are the same).

In order to elicit the components of our new metric, we computed the average scores of each of these metrics for each attribute type (i.e., gender, name, location, day, month and year) for both matches and mismatches over our entire dataset of over 16,000 pairs. We then computed the differences between these averages. Excepting binary discrimination for which the difference is always 1, we selected, for each attribute type, the metric with maximum difference. The results are summarized in Table 1.

Table 1: Metric Selection Table

Attribute Type	**Metric**
Gender	Binary Discrimination
Name	Soundex
Location	Jaro
Day	1-norm
Month	Dice
Year	1-norm

As a first attempt, our composite metric is a simple average of the selected metrics across the attributes, i.e.,

$$D(x, y) = \frac{\sum_i D_i(x_i, y_i)}{N} \qquad (1)$$

where N is the number of attributes with values in both x and y, D_i is the metric associated with the type of attribute i, and $D_i(x_i, y_i) = 0$ if either of x_i or y_i is missing.

Empirical results suggest that the heterogeneous metric is slightly superior to all of the homogeneous metrics. Figure 1 graphs the ROC curves parametrized by the similarity threshold (i.e., if distance exceeds threshold then the pairs are considered the same) for matching pairs.[1]

[1] ROC (Receiver Operating Characteristic) analysis has it origin in signal detection theory.

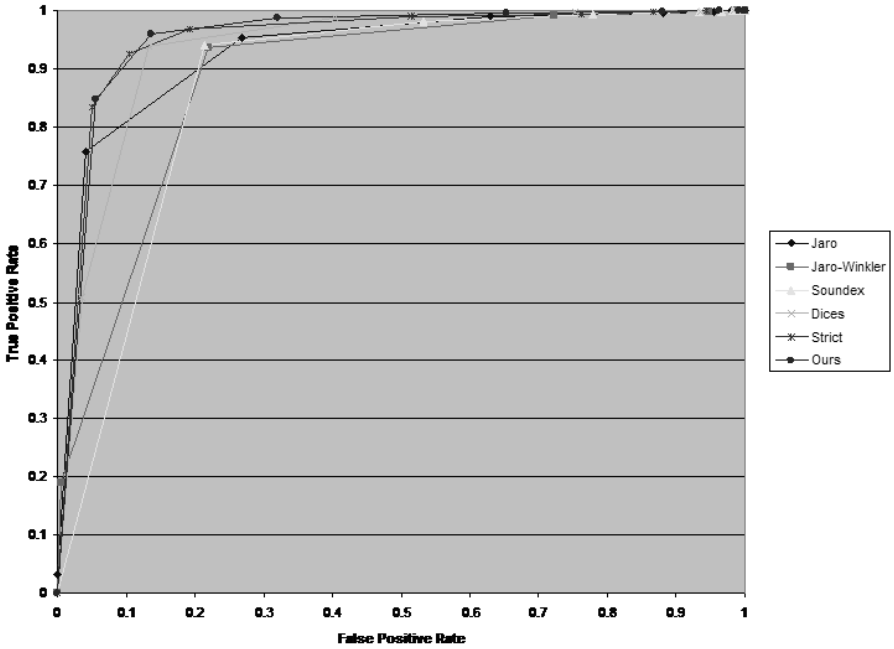

Figure 1: ROC Curves for Metrics

The possibility of weighing the various components of the heterogeneous metrics is the subject of further research, as is the use of Kolmogorov complexity-based similarity measures as a kind of parameter-free alternative (e.g., see [8, 1]).

3 Feature Selection

The Mining And Linking For Successful Information eXchange (MAL4:6) seeks to automatically learn which pedigree-based features are the most relevant in verifying pairs. MAL4:6 uses a large set of training data provided by the FHD and two scorecards to build a two-dimensional model of the graphical pedigree data.

The training examples are of the form $< id1 >< id2 >< Match(Y/N) >$. One of the scorecards is used for similar pairs, while the other is used for dissimilar pairs. The rows of the scorecards represent relationships (to the individuals of concern, e.g., self, father, child, grandmother) and the columns represent the

It was recently introduced in the field of Machine Learning/Data Mining [10]. ROC curves display true positive rate or sensitivity (i.e., ratio of number of matches found to total number of matches) versus false positive rate or 1 minus specificity (i.e., ratio of number of incorrectly found matches to total number of mismatches).

attributes of individuals (e.g., gender, first name, birth place, etc.), as depicted in Figure 2.

	Gender	1st Name	Bdate	Bplace	...
Self					...
Father					...
GrdMother					...
...	...	...	...	...	...

Figure 2: Sample Scorecard

Each cell then contains the similarity score of the individuals on the corresponding pedigree-based feature. The size of the search space (i.e., the number of relationships) is controlled by user-defined limits on the number of generations to consider. The scorecards are filled in according to the algorithm shown in Figure 3.

```
Input: id1, id2, match(Y/N), uplimit, downlimit
  Compute similarity score for attributes of id1 and id2
  Store in appropriate scorecard on row Self
  For each relative r between uplimit and downlimit
    Compute similarity score for attributes of id1.r and id2.r
    Store in appropriate scorecard on row labelled r
```

Figure 3: Scorecard-filling Algorithm

After it has gone over the training data, MAL4:6 uses the resulting scorecards to determine relevant features. A feature is classified as relevant if its similarity scores differ greatly between scorecards. A larger difference suggests that a feature will be helpful in determining the likelihood of two people matching. In our experiments, a feature is selected if the difference exceeds 0.5.

We ran experiments varying *uplimit* and *downlimit* from 0 to 4, comparing the average number of attributes used in the metric (see N in equation (1)). Figure 4 shows the ROC curves parametrized by the similarity threshold for matching pairs, for $uplimit = downlimit = 4$.

Overall, we observe that an average reduction of about one third in the number of attributes needed (from 64.2 to 19.8 in Figure 4) is possible without significant performance deterioration.

4 Graph-based Matching

Equiped with a metric and a feature selection mechanism, we assess the performance of MAL4:6 V0.1. The dataset is split into 2/3 training and 1/3 testing. The training data is used for feature selection (i.e., building the scorecards).

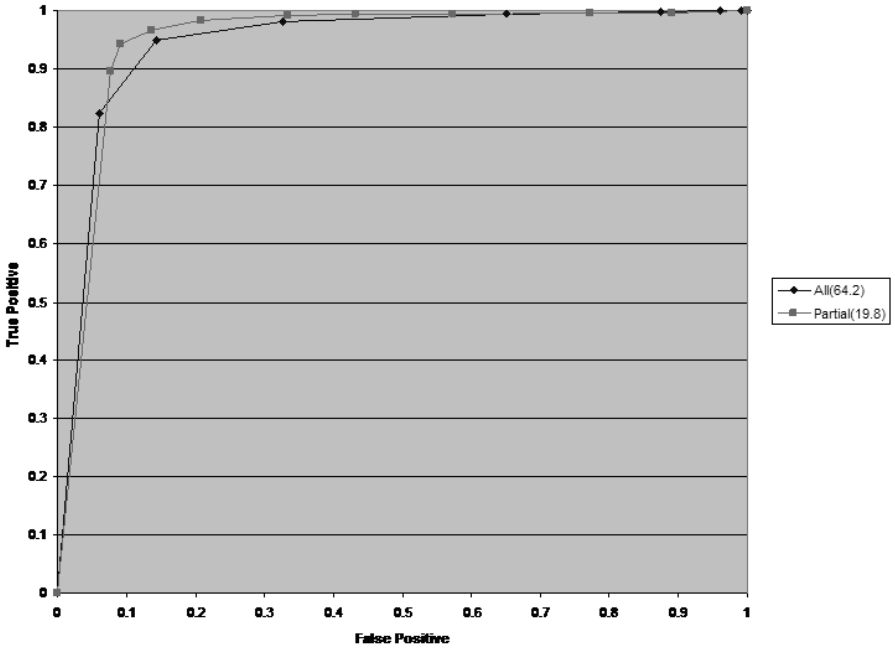

Figure 4: ROC Curves for Feature Selection (4-4)

We compute precision and recall on the test set using the heterogeneous metric (threshold set at 0.7), and compare them for the case when no pedigree data is used (i.e., only an individual's attributes are considered) and when pedigree-data is included. Results across values of *uplimit* and *downlimit* are fairly consistent. Table 2 shows the results when $uplimit = downlimit = 4$, for matching pairs (top 2 rows) and mismatching pairs (bottom 2 rows).

Table 2: Precision and Recall for Individual vs. 4 Generation

Individual-only	**4 Generations**
$R = 95.266$	$R = 94.617$
$P = 71.799$	$P = 71.766$
$R = 86.093$	$R = 86.169$
$P = 98.641$	$P = 98.358$

Although these early results do not show a significant difference between the graph-based approach and the standard individual-only approach, they nevertheless demonstrate promise as a number of enhancements are possible. In

particular, the current experiment assigns the same weights to all features regardless of their position in the pedigree chart. For example, similarity between grand-mothers' birth location has the same weight has similarity between individuals' last name. We expect to obtain better results by using generational weights, where the value of similarity decreases as the relationship becomes more distant. Alternatively, we are considering pairwise similarities across pedigrees rather than aggregate measures.

We also note that our dataset exhibits a ratio of 1:5 of matches to mismatches, which may explain the seemingly low absolute matching precision values. Many studies report precision on datasets where the ratio is held at 1:1, thus artificially boosting precision. We intend to further investigate this often ignored issue of designing systems under strong sample selection bias (e.g., see [14])

5 Conclusion

This paper introduces MAL4:6 V0.1, a pedigree-based record linkage approach. It discusses the value of using heterogeneous metrics for genealogical records and shows how features may be extracted across the pedigree chart whilst retaining performance. When applied to a large corpus of data, MAL4:6 V0.1 shows promise.

The process of using data mining to discover appropriate features to measure in a pedigree and explicitly using such graph-based features makes MAL4:6 different from existing record linkage methods. Most prior work has focused on improving methods of identifying string similarity and largely depends upon a single individual's attributes. Interest in graph/pedigree-based approaches is growing. For example, a limited use of family relationship information has been implemented in GENMERGE [4].

Future versions of MAL4:6 will focus on adding weights to the similarity metric and considering pairwise similarity across pedigrees. In addition, we will be investigating the issues associated with the skewed nature of the record linkage problem (i.e., 1:n ratio where n is large), as well as performing further experiments with additional data.

Acknowledgments

Data for our experiments was graciously provided by the Family History Department of the Church of Jesus Christ of Latter-day Saints. We express special thanks to Dallan Quass and Randy Wilson for insightful discussions on record linkage and encouragements with graph-based matching.

References

[1] Christen, P. and Goiser, K. (2005). Towards Automated Data Linkage and Deduplication. Submitted. (Available at http://datamining.anu.edu.au/publications/2005/pakdd2005-automated.pdf).

[2] Dice, L.R. (1945). Measures of the Amount of Ecologic Association Between Species. *Ecology*, **26**: 297-302.

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

[4] GENMERGE. http://www.genmerge.com/ (demonstrated at the *4th Annual Workshop on Technology for Family History and Genealogical Research*, Provo, UT.

[5] Giraud-Carrier, C. and Martinez, T. (1995). An Efficient Metric for Heterogeneous Inductive Learning Applications in the Attribute-Value Language. *Intelligent Systems (Proceedings of GWIC'94)*, E.A. Yfantis (Ed.), Kluwer Academic Publishers, Vol. 1, 341-350.

[6] Jaro, M. A. (1989). Advances in Record Linking Methodology as Applied to the 1985 Census of Tampa Florida. *Journal of the American Statistical Society*, **64**:1183-1210.

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

[8] Keogh, E., Lonardi, S. and Ratanmahatana, C.A. (2004). Towards Parameter-Free Data Mining. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 206-215.

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

[10] Provost, F. and Fawcett, T. (2001). Robust Classification for Imprecise Environments. *Machine Learning*, **42**:203-231.

[11] Quass, D. and Starkey, P. (2003). Record Linkage for Genealogical Databases. In *Proceedings of the Data Cleaning, Record Linkage and Object Consolidation Workshop*, Washington, D.C.

[12] Sorensen, T. (1948). A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species Content and its Application to Analyses of the Vegetation on Danish Commons. *K. Dan. Vidensk. Selsk. Biol. Skr.*, **5**: 1-34.

[13] Winkler, W. E. (1999). The State of Record Linkage and Current Research Problems. *Publication R99/04*, Statistics of Income Division, Internal Revenue Service. Available at http://www.census.gov/srd/www/byname.html.

[14] Zadrozny, B. (2004). Learning and Evaluating Classifiers under Sample Selection Bias. In *Proceedings of the Twenty-first International Conference on Machine Learning*, 903-910.

[15] Zobel, J. and Dart, P. (1995). Finding Approximate Matches in Large Lexicons. *Software–Practice and Experience*, Vol. 1, 331-345.