FRIL: A Tool for Comparative Record Linkage

Pawel Jurczyk, MS[1,2], James J. Lu, PhD[1], Li Xiong, PhD[1],
Janet D. Cragan, MD, MPH[2], Adolfo Correa, MD, MPH, PhD[2]
[1] Emory University, Mathematics and Computer Science, Atlanta GA;
[2] National Center on Birth Defects and Developmental Disabilities, Centers for Disease Control and Prevention, Atlanta GA

Abstract

A fine-grained record integration and linkage tool (FRIL) is presented. The tool extends traditional record linkage tools with a richer set of parameters. Users may systematically and iteratively explore the optimal combination of parameter values to enhance linking performance and accuracy. Results of linking a birth defects monitoring program and birth certificate data using FRIL show 99% precision and 95% recall rates when compared to results obtained through handcrafted algorithms, and the process took significantly less time to complete. Experience and experimental result suggest that FRIL has the potential to increase the accuracy of data linkage across all studies involving record linkage. In particular, FRIL will enable researchers to assess objectively the quality of linked data.

Introduction

The goal of record linkage is to find syntactically distinct data entries that refer to the same entity in two or more input files. The process is important for both data cleaning and integration in birth defects surveillance and research. Traditional interactive tools for record linkage provide users with a small number of parameters, consisting mostly of user-options for selecting similarity measures and decision models. In some cases, the user may also pick the search algorithm. The combination of choices typically does not provide sufficient granularities to produce results that are easily discernible. Hence for most research involving record linkage, the accuracy of the linked data is not well-understood, and often not discussed in the evaluation of the study.

As part of a surveillance program to monitor birth defects in the metropolitan Atlanta area, we have developed a fine-grained record integration and linkage tool (FRIL) to link a 12,700 record database from the Metropolitan Atlanta Congenital Defects Program (MACDP) with a 1.25 million record birth certificate database. The objectives of MACDP are to monitor births of infants with malformations for changes in incidence over time or patterns suggestive of environmental influences, to maintain a case registry for epidemiologic studies, to quantify the morbidity and mortality associated with birth defects, and to provide data for education and health policy decisions related to prevention[1]. Towards these objectives, MACDP conducts data linkages to enhance the completeness of birth defects surveillance data.

Background

The problem of record linkage is defined as follows. Given sets A and B of records, find a partition of A×B consisting of sets M (matched), U (unmatched), and P (possibly matched) that satisfy $M = \{(a, b) \mid a = b\}$ and $U = \{(a, b) \mid a \neq b\}$. A widely adopted record linkage approach is the probabilistic approach by Fellegi *et. al.*[2] First, a vector of similarity scores (or agreement values) is computed for each pair. Then, the pair is classified as either a match or non-match (or possibly matched) based on an aggregate of the similarity scores. Among methods used for classification we find rule-based methods that allow human experts to specify matching rules, unsupervised learning methods such as Expectation-Maximization (EM) that learns the weights or thresholds without relying on labeled data, and supervised learning methods that use labeled data to train a model, such as decision tree, naïve Bayesian or SVM. For detailed descriptions of those methods we refer readers to[3,4,5]. For computing similarities, various distance functions are used and studied. Complete descriptions of these methods can be found in[3,6,7], and several comparative evaluations of those methods have been performed[8,9].

FRIL adopts the probabilistic linkage approach. Its strength is the amount of control that the user has for tuning the accuracy and performance of linkages. In the remainder of the paper, we describe the full spectrum of user-tunable parameters available in FRIL and discuss their importance in the context of birth defect surveillance (BDS).

Methods

Among the user-controlled parameters in FRIL are certain algorithmic decision points that are usually

hidden in common linkage tools such as Link King[10], Link Plus[11] or LinkageWiz[12]. FRIL embodies the standard process of record linkage tools as described in, for example, TAILOR[13]. From the data sources, the user chooses a search method, a set of distance functions for measuring record similarity, and a decision model for accepting or rejecting a match. Iterative refinement of linkage is possible: unmatched records from one run of FRIL are available as input to a follow-up run using a different set of parameters. Graphical tools for reconciling schema discrepancy and for analyzing, validating and summarizing results have been incorporated. In addition, computerized learning tools are being developed to enable automatic parameters suggestion.

The workflow of FRIL is shown in Figure 1. The user specifies the initial input files. Each run involves the user specifying the search method, the distance function in the attribute comparison module and the decision model. Output consists of sets M, U, P and various summary statistics. Sets U and P may be fed back into FRIL using a different set of parameters.

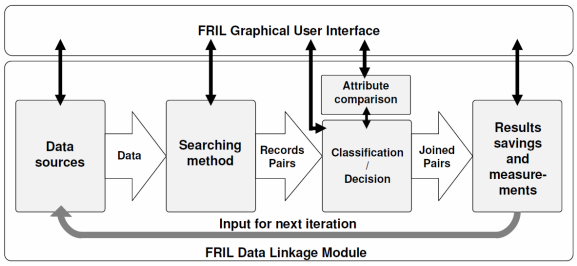

Figure 1: The FRIL architecture

Search methods. Search methods refer to algorithms for determining which pairs of records to compare between the data sources A and B and the attributes on which comparisons are made. FRIL implements two search methods: *nested loop join* (NLJ) and the *sorted neighborhood method* (SNM). NLJ performs an all to all comparison between *A* and *B* and is useful for small input data files.

The *sorted neighborhood method* (SNM) first sorts records of A and B over the relevant attributes, and follows by comparing only records within fixed windows ω_A and ω_B of records as these windows are advanced along the data files. Sorting moves records that have similar values (relative to the selected attributes) close together, presumably to within ω_A and ω_B of each other. This avoids the need to compare each record of one file against the entire data set of the second file. We call user inputs to ω_A and ω_B *window sizing* (**WS**).

An initial issue that the user must address when selecting attributes for comparison is to resolve possible discrepancies between schemas of *A* and *B*. Attributes are often labeled differently in the BDS study (e.g., "Baby Name" vs. "B Name"), and in some cases two attributes of one source map onto a single attribute of the other. FRIL allows users to identify attributes from one source to the other. Based on the user-specified mapping, FRIL will merge and split attributes or normalize attribute values on-the-fly if necessary. Attribute splitting is by separating data values based on regular expressions. We call this user input *attribute selection and mapping* (**ASM**). The resulting set of attributes to be used in the linking is denoted Φ.

When Φ contains more than a single attribute and searching is based on SNM, the choice of the dominant sorting attribute plays a critical role. In BDS, while the baby name attribute carries the greatest weight, using mother name as the dominant sorting attribute allows finding additional matches. The reason is that data discrepancy occurs more frequently on baby names, and, if used as the dominant sorting attributes, similar records may end up outside the tested window. We call the user input of the dominant sort attribute *sort ordering* (**SO**).

Distance function. The different string distance functions commonly used in record linkage are well understood[3]. In FRIL we have implemented edit distance, Soundex, Q-gram, and equality. All the functions have the same type: $a \times a \rightarrow [0,1]$, where a is an attribute in Φ. The smaller the function value, the closer to an exact match are the two inputs. FRIL allows users to choose a different distance function for each attribute in Φ. We refer to this input as a *distance functions selection* (**DFS**).

For each distance function f, the user is allowed to indicate the threshold for acceptance and rejection via a simple form of fuzzy logic. Specifically, if f_a is chosen as the distance function for attribute a, FRIL allows the user to specify the maximum $\max_{fa} \in [0,1]$ and the minimum $\min_{fa} \in [0,1]$ values for outright rejection and acceptance, respectively. Choosing 0 for min_{fa} implies that strict equality is required for acceptance. For values in between max_{fa} and min_{fa}, we use the membership function m_{fa}:

$$\begin{aligned} m_{f_a}(s_1, s_2) &= 1 \text{ if } f_a(s_1, s_2) \leq min_{f_a} \\ &= 0 \text{ if } f_a(s_1, s_2) > max_{f_a} \\ &= \frac{max_{f_a} - f_a(s_1, s_2)}{max_{f_a} - min_{f_a}} \text{ otherwise} \end{aligned}$$

We call this set of user inputs *attribute scoring* (**AS**). If $min_{fa} = 0$ and $max_{fa} = 1$, then the above function is the same as a continuous similarity function used in typical probabilistic linkage methods.

Decision model

Aside from selecting and mapping attributes (ASM), FRIL also allows a weight $\alpha_a \in [0,1]$ to be assigned to each attribute a in Φ. A higher weight reflects greater

importance. In BDS, matching baby name is more important than matching mother name or address. We refer to this user input *attribute weighting* (**AW**).

The final matching score for a pair of records r_1 and r_2 is the normalized weighted sum over all attributes:

$$score(r_1, r_2) = \frac{\sum_{a \in \Phi} \alpha_a m_{fa}(\pi_a(r_1), \pi_a(r_2))}{\sum_{a \in \Phi} \alpha_a}$$

Here π is the standard projection operator of the relational algebra. Again, the user may specify two weights, min_t and max_t to indicate the overall scores for match rejection and acceptance. Linked records with scores above max_t are considered matching, below min_t are unmatching, and in between are probable matches. A goodness of fit score is reported based on the following membership function:

$$\begin{aligned} M(r_1, r_2) &= 1 \text{ if } score(r_1, r_2) \geq max_t \\ &= 0 \text{ if } score(r_1, r_2) < min_t \\ &= \frac{score(r_1, r_2) - min_t}{max_t - min_t} \text{ otherwise} \end{aligned}$$

We refer to this user input the *record scoring* (**RS**). As an example, let $\Phi = \{a\}$, f_a be the edit distance, min_{fa}=0.5 and max_{fa}=1, α_a=1, min_t=0 and max_t=1. The following shows the scores and match results for three input record pairs (edit distance returns # edits as a fraction of the length of the longer string).

r_1	R_2	#edits	f_a	M
"AARON"	"ARON"	1	0.2	1.0
"AARON"	"ADAM"	4	0.8	0.4
"AARON"	"HUGH"	5	1	0.0

Observe that boolean join (or exact match) condition is a special case of the above discussion and may be obtained by choosing equality as the distance function, choosing $min_{fa} = max_{fa}$=0, and min_t=max_t=1. As finding correct weights and record scoring can be challenging, we are working to add unsupervised learning methods, such as EM, for suggesting good values. Table 1 includes a summary of the full space of parameters in FRIL.

Results and Discussion

An objective of our experimental evaluation is to present a process for obtaining the best possible linkage between two input data sources. The MACDP program is an active population-based surveillance system for birth defects that was established in 1967 by the CDC, Emory University, and the Georgia Mental Health Institute. The program has collected information on more than 12,700 cases of birth defects among the offspring of residents of the 5 central counties of Atlanta for years 1997-2006. As part of the surveillance program, a birth certificate database of 1.25 million records of children born in the state of Georgia for the same years was obtained from the Georgia Department of Human Resources.

The goal of linking the two data sets is to match each record from the MACDP database with a corresponding record in the birth certificates database. However, the two sources contain numerous metadata (i.e., schema level) and object-data heterogeneities. For example, the birth certificates database provides separate attributes for first and last name, and the same information is found under a single attribute in the MACDP database. The number of digits used for encoding year of birth also varies between the two sources. Metadata heterogeneities are resolved in FRIL through user specified attribute selection and mapping (ASM).

Parameter	Description	Possible values
WS	SNM window size selection	two integer numbers greater than 0
ASM	attribute selection and mapping	any subset combination of the data source attributes, including possible merging and splitting of attributes
AW	attribute weighting	real numbers between 0 and 1
SO	sort ordering	all permutations of the attributes
DFS	distance function selection	for each pair of attributes distance function from a set of available functions
AS	attribute scoring	two real numbers in the range [0,1] with respect to the distance function; 0 indicates identical values
RS	record scoring	two real numbers in the range [0,1] indicating acceptance and rejection thresholds of records with respect to the attribute scoring

Table 1: The FRIL Parameter Space

Object-data heterogeneity examples include mis-recording of information and missing data values. These are more difficult to handle and require the full range of FRIL features to resolve. The remainder of this section focuses on this type of heterogeneity.

Metrics. The two data sets in the BDS had been linked previously using a deterministic, rule-based approach. The results were obtained over 2-3 weeks through a combination of running the linkage program, adapted from a SAS program developed by the National Center for Health Statistics, and manual inspections. We use the fruit of this labor as the gold standard, G, against which our methods are compared. We evaluate using two standard metrics: precision and recall.

$$precision = \frac{\#\ of\ true\ positives}{\#\ of\ true\ positives + \#\ of\ false\ positives}$$

A true positive is a pair of correctly matched records, and a false positive is one that is incorrectly matched.

$$recall = \frac{\#\ of\ true\ positives}{\#\ of\ true\ positives + \#\ of\ false\ negatives}$$

For measuring improvements across experiments, precision and recall are better than sensitivity and specificity. Given the size of the data sets, the last

measure in particular is overshadowed by the large number of true negatives.

Data Characteristic. Among the 187 MACDP data attributes available for linking, Table 2 provides statistics about the data in some of the more intuitively important attributes in our data sources. In general, columns that have high number of *null* values are not good candidates to include in the join condition. Sources of *null* values vary, and may indicate (1) unknown, (2) non-applicable or (3) unrecorded parameters. Hence comparing attributes with *nulls* provide less information than comparing attribute with *non-nulls*. Initial choices of attributes are highlighted in bold italic font in the table.

Column name	**Percentage of not-null values**	
	MACDP data	Birth cert. data
***Birth date** (baby)*	*100%*	*100%*
***Name** (baby)*	*100%*	*100%*
Birth date (mother)	100%	100%
***Name** (mother)*	*100%*	*100%*
Birth date (father)	82%	83%
Name (father)	82%	83%
Hospital #	*100%*	*99.5%*
City	100%	97.5%
Zip code	*100%*	*99.8%*
Sex (child)	100%	100%

Table 2: Characteristics of columns in datasets

There are other attributes in the data sources not used in our experiments. The reason to work with a small subset of attributes is that it allows us to evaluate the ease of use and the utility of various features of FRIL without getting too involved in the details of the data.

Linkage experiments. With the attributes for the linking process fixed, we now describe experiments aimed at finding parameters of the join condition that produces the best linkage result. We focus on the SNM search method for its superior efficiency compared to NLJ. The six remaining parameters that must be decided are DFS, AS, AW, SO, RS, and WS.

Experiment 1. Our initial values for the DFS, AS and AW parameters are shown in Table 3. The top-down ordering of the attributes in the table corresponds to the parameter SO. Initial values for record scoring are min_t=max_t=0.61. In all experiments, we used ω_A=ω_B=8. For attributes with likely misspellings, the edit distance function is deployed with acceptance and rejection thresholds specified (as fraction of the length of the longer string) in the table. Results produced by the join condition as specified were quite good: precision at 95% and recall at 86%.

Experiment 2. We reviewed the false positives generated above and observed that non-matching date of birth was an important cause. We refined the join condition by increasing the weight assigned to the date of birth attribute to 0.35 and reduced the weight of the baby name attribute to 0.3. This resulted in a 98% precision and a slightly decreased recall of 85%.

Column name	**Metric**	**Weight**
Name (baby)	Edit dist. (min_{fa}=0.2, max_{fa}=0.25)	0.4
Birth date (baby)	Equality	0.25
Name (mother)	Edit dist. (min_{fa}=0.2, max_{fa}=0.25)	0.2
Zip code	Equality	0.1
Hospital #	Equality	0.05

Table 3: Initial join condition

Experiment 3. An examination of the remaining false positives showed a strong correlation to non-matching baby name, and in particular to the overly relaxed acceptance threshold for the edit distance function for baby name. We restricted the threshold (min_{fa}=0.15, max_{fa}=0.25), and it resulted in an improved precision value of 99% and no change in the recall.

Experiment 4. To address the relatively low recall rate we sifted through records that appeared in the gold standard *G* but that were not matched in Experiment 3. For most of these records, we observed that no attempted links were even made by FRIL. The reason lies in the sort ordering we used (again, indicated by the top-down ordering of attributes in Table 3). Using baby name as the dominant sorting attribute, two records with dissimilar values have the potential to occur far apart, beyond the window size, in the sorted files. However, significant mismatches in baby name often occur as the result of data entry conventions, e.g. for babies that have not been given a first name, the letter B is used to denote "Baby" (e.g., "Smith B"). In some cases, similar records appeared more than 1,200 records apart in the sorted file when sorted on baby name (Figure 2a). It turns out that while mother name is a semantically less significant attribute (i.e., carries less weight), it is a better dominant sorting attribute for many cases due to fewer variations in how its values are recorded. Figure 2b illustrates how the problem of Figure 2a is solved through a different sort ordering.

Rather than increasing the window sizes, which would hamper computational efficiency, we handled the problem with another feature of FRIL: the join summary. It allowed us to create an output of those MACDP records not joined in the initial run of the experiment (> 1500 records), and use them as input in a second-run of the experiment under a different set of parameters. By changing the dominant sorting attribute to mother name, the second run linked nearly 900 of the unmatched records from the first run. Thus the combined effect of the two runs yielded 99% precision and 95% recall. With a small window size of 8, each run of FRIL took approximately 20 minutes for the two data sources. The overall time for

completing the four experiments took less than two days. The remaining unmatched records have non-matching names, date of births, etc. Those records were joined manually in *G* with the assistance of human expertise. We also found 4 linkages that did not appear in *G*. This suggests another important utility of FRIL: that it can be used as a verification tool for existing linkage results. Table 4 shows a summary of the four experiments.

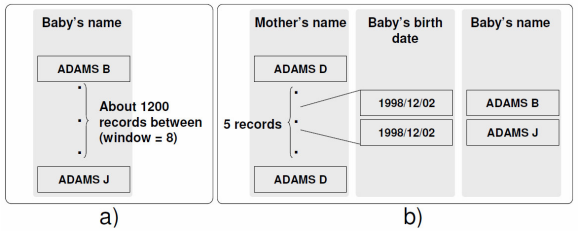

Figure 2: Impact of SO on compared records

	Experiment 1	Experiment 2	Experiment 3	Combined result
Precision	0.95	0.98	0.99	0.99
Recall	0.86	0.85	0.85	0.95

Table 4: Summary of results. Combined result contains linkages from two runs.

Conclusion and Ongoing Development

FRIL facilitates efficient and accurate record linkage over large data sources. The great flexibility of FRIL comes from the large number of fine-grained parameters that the user may tune, and it allowed us to link MACDP and birth certificates data efficiently and accurately (99% precision and 95% recall). By exploiting all the features of FRIL, we presented a process which enabled us to find good join condition. The benefits of FRIL extend beyond the results of linking. By revealing key algorithmic decision points for user inputs, the tool forces researchers to consider computational issues that impact accuracy and performance of the linkage process. As a result, researchers are able to judge the quality of the linked data scientifically and quantitatively. For already linked data, FRIL may also serve as a validation tool. Work on extending FRIL with several automated tools is ongoing. They include machine learning techniques to suggest values of certain parameters (e.g., attribute selection and weight). Borrowing query optimization techniques from databases, window size and sort ordering may also be suggested. We are optimistic that FRIL will facilitate many future projects based on birth defects surveillance data and other public health surveillance projects.

References

1. A. Correa, J.D. Cragan, M.E. Kucik, C.J. Alverson, S.M. Gilboa, R. Balakrishnan, M.J. Strickland, C.W. Duke, L.A. O'Leary, T. Riehle-Colarusso, C. Siffel, D. Gambrell, D. Thompson, M. Atkinson, J. Chitra. Metropolitan Atlanta Congenital Defects Program 40th Anniversary Edition Surveillance Report. Birth Defects Research Part A: Clinical and Molecular Teratology 79(2): 65-186, 2007.
2. I. P. Fellegi and A.B. Sunter. A Theory for Record Linkage. JASA, 64(328): 1183-1210, 1969.
3. A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios. Duplicate record detection: A survey. IEEE TKDE, 19(1):1–16, 2007.
4. A. Halevy, A. Rajaraman, and J. Ordille. Data integration: the teenage years. In Proc. of the VLDB, 2006.
5. W. Winkler. Overview of record linkage and current research directions. U.S. Census Bureau, Technical Report, 2006.
6. G. Navarro. A guided tour to approximate string matching. ACM Computing. Survey, 33(1):31–88, 2001.
7. E. S. Ristad and P. N. Yianilos. Learning string edit distance. Technical Report CS-TR-532-96, Department of Computer Science, Princeton University, 1996.
8. W. Cohen, P. Ravikumar, and S. Fienberg. A comparison of string metrics for matching names and records. In Proc. Of KDD Workshop on Data Cleaning, 2003.
9. E. Porter and W.Winkler. Approximate string comparison and its effect on an advanced record linkage system. U.S. Bureau of the Census, Research Report, 1997.
10. K.M. Campbell. Rule Your Data with The Link King (a SAS/AF application for record linkage and unduplication) . SUGI 30, 2005.
11. K.K. Thoburn, D. Gu and T. Rawson. Link Plus: Probabilistic Record Linkage Software. 2nd Probabilistic Record Linkage Conference Call, 2007.
12. Record linkage software. Version 5.0. LinkageWiz Inc. Available from http://www.linkagewiz.com/.
13. M. G. Elfeky, V. S. Verykios, and A. K. Elmagarmid. TAILOR: A record linkage toolbox. In Proc. of the ICDE, 2002.