An Empiric Weight Computation for Record Linkage Using Linearly Combined Fields' Similarity Scores

Xinran Li, Aline Guttmann, Jacques Demongeot, Jean-Yves Boire and Lemlih Ouchchane

Abstract - **Record linkage is the task of identifying which records from one or more data sources refer to the same entity. Many record linkage methods were introduced and applied over the last decades. In general, the principle is to compare a range of available identifier fields in record pairs among different data sources, in order to make a linkage decision. The Fellegi-Sunter probabilistic record linkage (PRL-FS) is one of the most commonly used methods. To obtain a better performance, Winkler proposed an enhanced PRL-FS method (PRL-W) that takes into account field similarity, but its implementation requires the estimation of much more parameters which complicates the task. Consequently, we propose to develop a method that contains the best features in the PRL-FS and the PRL-W methods: simplicity of parameters estimation and consideration of fields' similarities. We hypothesize that our record linkage method outperforms the PRL-FS, and can achieve a similar performance of the PRL-W. This paper presents briefly the PRL-FS and PRL-W methods, and describes in details how to combine fields' similarity scores to create a novel record pair weight. Simulated data sets were used to assess and to compare these three methods regarding their ability to reduce the rates of false matches and false non-matches.**

I. Introduction

To obtain a more complete patient's medical history whether in the context of care delivery or epidemiological studies, correctly and efficiently linking the same patient's data is crucial. Usually, different healthcare databases do not share a unique patient identifier, except the social security number. But the use of this number to link patient's data is not allowed in many countries because of privacy protection legislation [1]. In such a case, we can compare a range of corresponding identifier fields (e.g. first name, last name, birth date and sex) in record pairs among different databases, and classify each compared record pair as a match or a non-match according to all or a part of its fields' agreement/disagreement. But some fields provide "more information more reliably" than others [2], and these fields are possibly subject to misspellings and typographical errors [3], [4]. Therefore, we need an efficient record linkage method taking into account all these factors.

X. L., A. G., JY. B. and L. O. are with ISIT, UMR CNRS UdA 6284, Auvergne University, F-63001, France (corresponding author e-mail: xinran.li1@udamail.fr; authors e-mails: aline.guttmann@udamail.fr; j-yves.boire@udamail.fr; lemlih.ouchchane@udamail.fr).

J. D. is with AGIM, FRE CNRS 3405, Joseph Fourier University, La Tronche University School of Medicine, Grenoble, F-38700, France (e-mail: jacques.demongeot@agim.eu)

The PRL-FS is a commonly used probabilistic record linkage method formalized by Fellegi and Sunter [5]. In this method, each corresponding field of record pairs is assigned an agreement weight and a disagreement weight based on log likelihood ratios [6]. For each record pair, a composite weight is computed by summing each field's agreement or disagreement weight. When a field **agrees** (the contents in field to compare are the same), the field agreement weight is used for computing the composite weight; otherwise the field disagreement weight is used. A record pair with a composite weight above a certain threshold value is classified as a match, while a record pair with a composite weight below a certain threshold value is classified as a non-match [7].

While the PRL-FS method is relatively easy to implement, it has a drawback that each field has only two possible weights: agreement weight if the field exactly agrees and disagreement weight otherwise. For example, in two record pairs, the field *first name* such as (*Sebastian, Sebastien*) and (*Sebastian, Joe*) are assigned the same weight as both have a disagreement between two strings. But the first pair is much more likely to refer to the same person than the second pair, because the misspellings and typographical errors in fields like first name and last name could be reached in more than 30% of records [3], [4], [6]. Consequently, Winkler proposed the PRL-W method that takes into account **field similarity** (similarity of two strings for a field within a record pair) in the calculation of field weights, and proved its outperformance over the PRL-FS [8].

In the PRL-W method, a similarity score is calculated for the string pair in each corresponding field. The higher the score, the more similar the string pair, and this score is standardized between 0 (no similarity) and 1 (exact agreement). The closed interval [0,1] is then partitioned into a collection of disjoint subintervals, each compared field is assigned an approximate agreement weight depending on which subinterval the score falls in [8]. This method can provide more information about record pairs for discriminating between them in the match/non-match classification. Therefore, the introduction of approximate agreement weights results in much more parameters to estimate. Using the subinterval length chosen by Winkler, implementation of this method will require the estimation of 42 parameters per string field, unlike the PRL-FS where only 2 parameters per field are needed to estimate.

Consequently, we propose to develop a record linkage method where the record pair weight is computed by a linear combination of fields' similarity scores used in PRL-W method. For the coefficients in this linear combination, we

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

choose empirically the fields' agreement weights used in PRL-FS method, which reflect the importance of information provided by each field. We hypothesize that the record linkage method using linearly combined fields' similarity scores (RL-CS) outperforms the PRL-FS, and can achieve a similar performance of the PRL-W. Using simulated data sets, we implemented the RL-CS, the PRL-FS and the PRL-W methods. Then, we evaluated and compared these three methods in their ability to reduce the rates of **false matches** (record pair classified as a match for different persons) and **false non-matches** (record pair classified as a non-match for the same person).

II. Material and Methods

A. Simulated Data Sets

To evaluate the performance of record linkage methods, which could be applied to data sets with different sizes and data qualities, we have to know: (1) the **truth of matches** (record pair belongs to the same person or to different persons in reality) to which we can compare our **linkage decision** (record pair is classified as a match or a non-match), and (2) the proportion and type of errors in each data sets. Such work using real data would require extremely costly verifications without being certain to find all the false linkage decisions and errors in data sets. Therefore, we chose to use simulated data sets to perform our study.

Fig. 1 shows the key steps for creating our simulated data sets. We first generated a sample of N_E fictitious records. Each record consists of five fields: first name, last name, birth date, sex and a unique identifier of record among all data sets. From these generated records, N_A and N_B records (with $N_A + N_B > N_E$) were randomly drawn without replacement to constitute data sets A and B, respectively. Errors were then introduced into a proportion of randomly selected records of data sets A and B (errors introduction involves only the first four fields in record). Both the type and proportion of errors can be modulated to construct different data sets that could be possibly encountered in real linkage situations.

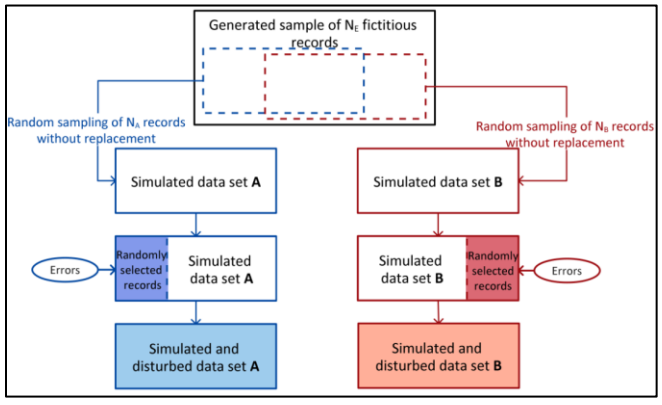

Figure 1. Simulation of data sets

The types of errors introduced in our simulated data sets were omission, insertion, substitution or transposition of one or more characters or numbers among a string in field, these types are the most common spelling errors in identifier fields according to a patient data validation study [3]. In a data set, these errors will occur in a certain proportion of records. In the literature, this proportion is ranged from about 8.5% to 36.5% in different investigated data sets [1], [3], [9]. Therefore, we simulated different data sets, that the error introductions were applied to approximately 10%, 20% and 30% of randomly selected records in data set. For the distribution of errors within these corrupted records, there were 32% errors in first name, last name and birth date respectively, and 4% errors in sex. A record can have more than one field having errors. In our simulated date sets, about 90% records had errors in 1 field, 9% records had errors in 2 fields, 0.9% records had errors in 3 fields and 0.1% records had errors in 4 fields.

For each configuration (a specified size and error rate for data set), the simulation was repeated 100 times, so the evaluation of record linkage methods presented for each configuration reflects the (average) result of 100 simulations.

B. Field Similarity Score Computation

The similarity score used in the PRL-W method is the Jaro-Winkler distance [8], which is a measure of similarity between two strings (contents in a corresponding field within a record pair). The similarity score for two strings is computed by taking simultaneously into account the length of each string, the number of common characters and the number of character transpositions in the two strings [10]. This score is normalized between 0 and 1, with 0 reflecting the absence of similarity and 1 an exact agreement. We used the string comparison function "*jarowinkler*" in the R package "*RecordLinkage*" for computing the Jaro-Winkler similarity score (JWSS) [11], [12]. For example, using the above function with default arguments, the JWSS for first names pairs (*Sebastian, Sebastien*) and (*Sebastian, Joe*) are 0.9556 and 0.4815, respectively. For the field date of birth, we considered its date format values as strings for the similarity score computation, the JWSS for date of birth can therefore be computed as for names. For the field sex, its values are standardized on "*M*" or "*F*" with a data pre-processing, so that the JWSS for sex can only be 0 or 1.

To compute the record pair weight in the RL-CS method, the JWSS of each field will be linearly combined, and for the coefficients in this linear combination, we proposed to use the fields' agreement weights in the PRL-FS method.

C. Field Agreement Weight in the Fellegi-Sunter Method

The field agreement weight used in the PRL-FS method can represent the importance of information provided by field. For example, if there is no error in fields, two persons with the same last name are much more likely to refer to the same person than two persons with the same sex. Therefore, the last name agreement weight should be much higher than the sex agreement weight. The field agreement weight is a log likelihood ratio based on the *m* and *u* probabilities, where *m* is the probability that a field agrees given that the record pair belongs to the same person, and *u* is the probability that a field agrees given that the record pair belongs to different persons [13]. For a given record pair, if field *i* agrees, then the weight for this field is

$$w_i = log_2(m_i/u_i) \quad (1)$$

The estimation of *m* and *u* probabilities can be performed with the expectation maximization (EM) algorithm.

D. Parameters Estimation

Using the simulated data sets where the truth of matches is known, it would be straightforward to compute the parameters m_i and u_i as:

$$m_i = \frac{\text{\#record pairs involving the same person where field } i \text{ agrees}}{\text{\#record pairs involving the same person}}$$

$$u_i = \frac{\text{\#record pairs involving different persons where field } i \text{ agrees}}{\text{\#record pairs involving different persons}}$$

In addition, we also need to compute the parameter p (the proportion of pairs involving the same person within all possible record pairs of the two data sets), which can be used to establish the decision threshold to classify record pairs as matches or non-matches [7].

However, in practice, the truth of matches is unknown. The EM algorithm -a method for finding maximum likelihood estimates of parameters in probabilistic models with unobserved variables- can thus be used to estimate the parameters *m, u* and *p*. This method starts with a reasonable initial guess of the parameters. The E step is to calculate the expectation of the likelihood function using the current parameters. The M step is to maximize the likelihood function using the expected value computed on the E step to obtain new parameters. We iterate the E and M steps until the parameters converge [14]–[16].

The data log-likelihood is [17]:

$$\ln f(m,u,p|g,\gamma) = \sum_{j=1}^{N} g_j \ln\left(p\prod_{i=1}^{n} m_i^{\gamma_i^j}(1-m_i)^{1-\gamma_i^j}\right) + \sum_{j=1}^{N}(1-g_j)\ln\left((1-p)\prod_{i=1}^{n} u_i^{\gamma_i^j}(1-u_i)^{1-\gamma_i^j}\right) \quad (2)$$

where:

g_j = unobserved value indicating whether or not the record pair *j* belongs to the same person (this value should be 1 if the record pair *j* belongs to the same person; 0 otherwise)

γ_i^j = observed agreement or disagreement of the field *i* in the record pair *j* (1 for agreement, 0 for disagreement)

N = total number of record pairs

n = total number of fields

For the E step, the expectation of g_j is computed as follows [17]:

$$E[g_j|\gamma] = \frac{p\prod_{i=1}^{n} m_i^{\gamma_i^j}(1-m_i)^{1-\gamma_i^j}}{p\prod_{i=1}^{n} m_i^{\gamma_i^j}(1-m_i)^{1-\gamma_i^j} + (1-p)\prod_{i=1}^{n} u_i^{\gamma_i^j}(1-u_i)^{1-\gamma_i^j}} \quad (3)$$

For the M step, by equating the partial derivatives (for each of the parameters *m*, *u* and *p*) of the function (3) to zero, the unknown parameters can be estimated [17]:

$$\hat{m}_i = \frac{\sum_{j=1}^{N} g_j\gamma_i^j}{\sum_{j=1}^{N} g_j} \quad (4) \qquad \hat{u}_i = \frac{\sum_{j=1}^{N}(1-g_j)\gamma_i^j}{\sum_{j=1}^{N}(1-g_j)} \quad (5) \qquad \hat{p} = \frac{\sum_{j=1}^{N} g_j}{N} \quad (6)$$

In our study, initial values for the parameters m_i, u_i and p were 0.5, 0.1 and 0.0001, respectively; convergence criterion was the difference between values of estimated parameters in two consecutive iterations less than 10^{-8}. The m_i and u_i estimated in the last iteration will be used to compute the field weight w_i.

E. Record Pair Weight Computation

We proposed to compute the weight for record pair *j* as follows:

$$w^j = \sum_{i=1}^{n} w_i \times JWSS_i^j \quad (7)$$

In this weight computation, the importance of information provided by each field (w_i) is constant, which is determined by (1). For each record pair, its fields' JWSS are calculated, we denote the JWSS of field *i* within record pair *j* as $JWSS_i^j$. The weight for record pair *j* represents a linear combination of $JWSS_i^j$.

F. Record Pair Classification

In a record linkage process for data sets A and B, we compared each record from data set A with all records in the data set B, so that $N_A \times N_B$ record pairs were compared, and each of them is assigned a weight (w^j). Record pairs were then classified as matches or non-matches according to a given decision threshold (record pairs with weights above the threshold are considered as matches; record pairs with weight below the threshold are considered as non-matches).

Basing on the truth of matches, the optimal decision threshold is easy to find using a ROC-analysis [18]. As this information is unknown in practice, we used the estimated value of p to choose a decision threshold. The number of pairs involving the same person in all possible record pairs between data sets A and B can be calculated by $p \times N_A \times N_B$. Sorting record pairs according to their weight (w^j) into ascending order, it seems reasonable to choose a decision threshold value near the weight of the $((1-p) \times N_A \times N_B)^{th}$ ordered record pair.

G. Evaluation and comparison of record linkage methods

With the same data sets A and B, we performed record linkage process by using each of PRL-FS, PRL-W and RL-CS methods. The implementation of the PRL-FS method is frequently presented in the literature [17], [19]–[21]. We described in details how to implement the PRL-W method in a previous paper [22]. The RL-CS was implemented using record pair weight w^j and estimated p. Finally, we compared these methods in their ability to reduce the number of falsely classified record pairs.

In our study, the comparisons were performed using data sets with different configurations (size and proportion of errors for data set). We have chosen the following configurations in data sets A and B: number of records equal to 500×2^n $(n = 0, 1, 2)$; errors in 10%, 20% and 30% of records in each data set. For each configuration, the simulations of data sets A and B were repeated 100 times.

III. Results

The following results were for linkage scenario N_A=1000, N_B=1000 and errors in 30% of records. Fig. 2 shows the numbers of falsely classified record pairs (out of 10^6 compared record pairs) in 100 "runs", carried out by the three record linkage methods. We can observe that the RL-CS and the PRL-W had a similar performance, and both of them outperformed the PRL-FS.

Using PRL-FS, PRL-W and RL-CS methods, they produced 71.8, 12.1 and 11.8 falsely classified record pairs in average in 100 simulations, and their standard deviations were 16.5, 8.7 and 6.8, respectively.

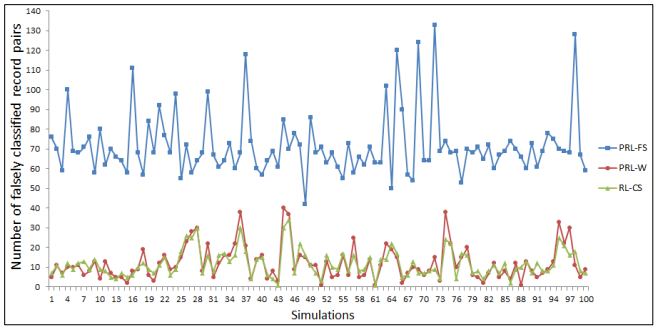

Figure 2. Numbers of falsely classified record pairs

In addition, we compared the computational time of these three methods. We used the computer having a CPU of Intel(R) Xeon(R) E5-2620, 2.0 GHz, and 16 GB RAM. The computational times of PRL-FS, PRL-W and RL-CS methods were 34.3, 75.6 and 36.6 minutes in average in 100 simulations, respectively. Among the three methods, the PL-CS is the most efficient and stable method to reduce falsely classified record pairs.

IV. Discussion and Conclusion

Using simulated data sets with different configurations, we have proven our hypothesis that the RL-CS outperforms the PRL-FS and can achieve a similar performance to the PRL-W. The PRL-W is based on a strong theoretical foundation, but we preferred the RL-CS because of the simplicity of its implementation: 2 parameters to estimate per field instead of 42 in the PRL-W. The RL-CS has also the advantage to take into account both the importance of information provided by each field and the field similarity.

In our study, the parameters (m, u and p) used in each method were obtained by estimations. The performances of these record linkage methods could be dependent on estimator's performance. We compared therefore the results of the linkages using estimated and observed parameters. In each linkage, the use of estimated and observed m and u (for weight computation) led to the same numbers of false matches and false non-matches; the use of observed p (for threshold choice) led in general to more accurate decisions.

This study illustrates the linear combination of fields' JWSS for creating a novel record pair weight, and demonstrates the outperformance of the RL-CS method by using simulated data sets. The data simulation is performed according to some common spelling errors and the proportion of errors in databases presented in the literature, but the simulated error types are far from exhaustive. For example, the inversion between first and last name or between maiden and married name might be common errors; the problem of missing data in fields also occurs frequently [23], which we should integrate into our future developments. Furthermore, as we cannot anticipate all possible scenarios in a real linkage work, a sampling method defining the types and rates of error in the current data sets would be highly desirable to adapt and improve the RL-CS to a specific context.

References

[1] M. Tromp, A. C. Ravelli, G. J. Bonsel, A. Hasman, and J. B. Reitsma, "Results from simulated data sets: probabilistic record linkage outperforms deterministic record linkage," *J. Clin. Epidemiol.*, vol. 64, no. 5, pp. 565–572, mai 2011.

[2] M. A. Jaro, "Probabilistic linkage of large public health data files," *Stat. Med.*, vol. 14, no. 5–7, pp. 491–498, 1995.

[3] C. Friedman and R. Sideli, "Tolerating spelling errors during patient validation," *Comput. Biomed. Res.*, vol. 25, no. 5, pp. 486–509, Oct. 1992.

[4] E. H. Porter and W. E. Winkler, "Approximate string comparison and its effect on an advanced record linkage system," in *Advanced Record Linkage System. US Bureau of the Census, Research Report*, 1997.

[5] I. P. Fellegi and A. B. Sunter, "A theory for record linkage," *J. Am. Stat. Assoc.*, vol. 64, no. 328, pp. 1183–1210, 1969.

[6] S. L. DuVall, R. A. Kerber, and A. Thomas, "Extending the Fellegi-Sunter probabilistic record linkage method for approximate field comparators," *J. Biomed. Inform.*, vol. 43, no. 1, pp. 24–30, Feb. 2010.

[7] V. J. Zhu, M. J. Overhage, J. Egg, S. M. Downs, and S. J. Grannis, "An empiric modification to the probabilistic record linkage algorithm using frequency-based weight scaling," *J. Am. Med. Inform. Assoc. JAMIA*, vol. 16, no. 5, pp. 738–745, Oct. 2009.

[8] W. E. Winkler, "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage.," 1990.

[9] W. E. Winkler, "Approximate string comparator search strategies for very large administrative lists," *Statistics*, p. 02, 2005.

[10] W. E. Winkler, "Overview of record linkage and current research directions," BUREAU OF THE CENSUS, 2006.

[11] M. Sariyar and A. Borg, "The RecordLinkage Package: Detecting Errors in Data."

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

[13] T. Blakely and C. Salmond, "Probabilistic record linkage and a method to calculate the positive predictive value," *Int. J. Epidemiol.*, vol. 31, no. 6, pp. 1246–1252, 2002.

[14] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Stat. Soc. Ser. B Methodol.*, pp. 1–38, 1977.

[15] T. K. Moon, "The expectation-maximization algorithm," *Signal Process. Mag. IEEE*, vol. 13, no. 6, pp. 47–60, 1996.

[16] C. B. Do and S. Batzoglou, "What is the expectation maximization algorithm?," *Nat. Biotechnol.*, vol. 26, no. 8, pp. 897–899, août 2008.

[17] M. A. Jaro, "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida," *J. Am. Stat. Assoc.*, vol. 84, no. 406, p. 414, Jun. 1989.

[18] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer, 2009.

[19] W. E. Winkler, "Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage," in *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 2000, pp. 667–671.

[20] S. J. Grannis, J. M. Overhage, S. Hui, and C. J. McDonald, "Analysis of a Probabilistic Record Linkage Technique without Human Review," *AMIA. Annu. Symp. Proc.*, vol. 2003, p. 259, 2003.

[21] C. Samuels, "Using the EM Algorithm to Estimate the Parameters of the Fellegi-Sunter Model for Data Linking." Methodology Advisory Committee Papers, Australian Bureau of Statistics, 2012.

[22] X. Li, A. Guttmann, S. Cipiere, L. Maigne, J.-Y. Boire, and L. Ouchchane, "Implementation of an Extended Fellegi-Sunter Probabilistic Record Linkage Method Using the Jaro-Winkler String Comparator." [Online]. Available: http://emb.citengine.com/event/bhi-2014/paper-details?pdID=122. [Accessed: 06-Jun-2014].

[23] T. C. Ong, M. V. Mannino, L. M. Schilling, and M. G. Kahn, "Improving record linkage performance in the presence of missing linkage data," *J. Biomed. Inform.*, Feb. 2014.