Advanced Topics in Information Systems B

A Survey of Probabilistic Record Matching Models, Techniques and Tools

Federico Maggi

Cycle XXII
Scientific Report TR-2008-22

Abstract

Probabilistic record linkage regards the use of *stochastic* decision models to solve the problem of *record linkage* (also known as record matching). Data quality has became a key aspect in many institutions and the demand for novel, effective techniques is increasing. Record linkage in general has been studied in the last three decades and a solid probabilistic decision framework has been proposed along with several extensions and specific estimation methods. This paper is a survey work narrowed to the most recent and promising approaches also including a selection of data cleansing tools based on probabilistic decision models.

1 Introduction

Heterogeneous and distributed data sources are often populated, manipulated and deployed by several different agents or companies all over the world. It is indeed vital to have effective methods for dealing with errors. The assessment of data quality is indeed becoming a key process, also for modern enterprises and small Web service providers, and not only for the management of large legacy databases of census or health agencies. In such a scenario, the research community and the industry need for novel contributions in terms of fast, optimized, and accurate data analysis and *cleansing* techniques.

Missing fields, records or integrity constraints, inconsistencies between tables, duplicated data are examples of "symptoms" indicating poor or insufficient data quality. Generally speaking, poor data quality is related to *heterogeneity* in the data. Heterogeneity can be either *structural* or *semantic*; *structural heterogeneity* occurs when the data is inconsistent within the scope of the information systems under consideration, which in practice means that *different* sources may store the (same) data using *different*

representations (e.g., different fields, different types, slightly different field values, typography errors, etc.). *Semantic heterogeneity* occurs when the same, similar structure is used to represent entities that are intrinsically different to each others. The latter is subject of this paper; more precisely, we focus on one specific issue that can arise when semantic inconsistencies occur: *data duplication*.

Duplicated records are obviously unwanted; in order to remove them the core problem is how to detect similar records. Such a problem has been called *record linkage* [Fellegi and Sunter, 1969], or *record matching*, and it is the task of accurately label records pairs corresponding to the same entity from different sources. In other words, the goal of a record linkage algorithm is to identify records that do not match completely. This paper provides an expository analysis on the use of probabilistic decision models for record linkage. In particular, we provide an overview of the basic theory of probabilistic linkage [Fellegi and Sunter, 1969]. Such a theory is abstract and rigorous thus it has been the main reference for the development of several proposals. Our goal is to survey and compare both the classical methods and the more innovative techniques in the literature of the last three decades.

The remainder of this paper is structured as follows. Narrowing the focus on probabilistic methods, in Section 2 the problem of record linkage is stated and the basic definitions are given as well as the notation that will be used; this section also contains the main taxonomic dimensions utilized to classify record linkage algorithms. Section 3 introduces how the Bayesian decision model has been applied to the problem of record linkage, including a detailed comparison of the most relevant variants proposed so far in the literature. A descriptive list of the available data cleansing and record matching tools is provided in 4 with particular attention to two of the most recent open source applications.

2 Record Linkage and Probabilistic Matching: Notation and Basic Definitions

The most rigorous mathematical framework to formalize the record linkage problem has been proposed by [Fellegi and Sunter, 1969] as an extension of the early work of [Newcombe and Kennedy, 1962]. In this paper, we draw the notation and the concepts defined in the theory of *Fellegi and Sunter* (FS) to present the reviews of the selected approaches.

Record linkage algorithms work on two sets (or files) of records denoted as A and B; we will use lowercase letters to indicate records belonging to each set, $a \in A$ and $b \in B$. The available information regarding one record is denoted with $\alpha(a)$ and $\beta(b)$, respectively.

The comparison set $A \times B$ is partitioned into the two subsets

$$M = \{(a, b) \in A \times B \mid a = b\}$$

of *matching* pairs and

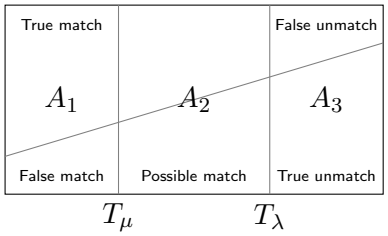

Figure 1: Partitioning of the decision space.

$$U = \{(a, b) \in A \times B \mid a \neq b\}$$

of *unmatching* pairs. Note that $M \cup U = A \times B$ and $M \cap U = \varnothing$.

The two data-sets are compared by means of a *comparison vector* γ which is a vector function of the records pairs $(a, b) \in A \times B$ (or, more precisely, $(\alpha(a), \beta(b))$). The space of all possible values of γ is Γ:

$$\gamma^j = [\gamma_1, \gamma_2, \ldots, \gamma_i, \ldots, \gamma_K]^T$$

with $\gamma_i \in \{0, 1\}$. In particular, $\gamma_j^i = 0$ if field i of the agrees (i.e., matches w.r.t. the criterion specified by γ) for pair j, that is $(\alpha(a), \beta(b))$.

From a probabilistic point of view, γ is an event thus a conditional probability can be attached to the vector:

$$m(\gamma) = P(\gamma \mid (a, b) \in M) = P(\gamma \mid M)$$

$$u(\gamma) = P(\gamma \mid (a, b) \in U) = P(\gamma \mid U)$$

that are, respectively, the probability of observing γ given a "match" and the probability of observing γ given a "non match".

A record linkage algorithm labels record pairs as A_1, if they match, A_3, if they do not match, and A_2 if they possibly match. Hence, a *linkage rule* is a decision function $d(\gamma) : \Gamma \mapsto D$ which assigns probabilities to each of the there decisions:

$$d(\gamma) = \{P(A_1 \mid \gamma), P(A_3 \mid \gamma), P(A_2 \mid \gamma)\}$$

and it is such that $P(A_1 \mid \gamma) + P(A_3 \mid \gamma) + P(A_2 \mid \gamma) = 1$. D denotes the set of all possible decisions.

As for every classification algorithm, there are two kind of errors: type I errors, and type II errors. The first refers to a *false match*, denoted by

$$\mu = P(A_1|U) = \sum_{\gamma \in \Gamma} u(\gamma) P(A_1|\gamma)$$

while the second refers to a *false non-match* and it is denoted by

$$\lambda = P(A_3|M) = \sum_{\gamma \in \Gamma} m(\gamma) P(A_3|\gamma).$$

Without going into the details, given μ and λ, the theory of FS proves that there exist an optimal linkage rule (i.e., an optimal decision function $d(\gamma)$). Such a rule minimizes the amount of records requiring manual review, that is the probability of labeling a record with A_2. The decision function relies onto two thresholds T_μ, T_λ which partition the decision space into the three areas A_1, A_2, A_3. Given the likelihood ratio $l(\gamma) = \frac{m(\gamma)}{u(\gamma)}$, if $\mu = \sum_{i=1}^{n} u(\gamma_i)$ and $\lambda = \sum_{i=n'}^{|\Gamma|} m(\gamma_i)$, with $n < n'$, the optimal thresholds are $T_\mu = l(\gamma_n)$ and $T_\lambda = l(\gamma_{n'})$. Figure 1 gives a graphical view of of how the decision space is partitioned.

To make the algorithm feasible in practice, Fellegi and Sunter state that if the components of γ are assumed to be mutually statistically independent w.r.t. each of the conditional distributions, the likelihood function can be rewritten as logarithms as:

$$w^k(\gamma_k) = \log m(\gamma^k) - \log u(\gamma^k)$$

to obtain a suitable test statistic

$$w(\gamma) = w^1 + w^2 + \cdots + w^K$$

and $w^1 + w^2 + \cdots + w^K$ are called *weights*. The different linkage techniques differs in the specific algorithm used to compute weights, for instance the one described in Section 3.1.4.

Finally, it must be noticed that for agreement configurations $m(\gamma^k)$ tends to one and $u(\gamma^k)$ tends to zero; on the other hand, for disagreement configurations $m(\gamma^k)$ is close to zero while $u(\gamma^k)$ is close to one.

2.1 Classical Hypothesis Test

For the sake of clarity, the above defined decision framework can be reformulated using the theory classical hypothesis test. In particular, if α is the test significance and $\mathcal{G}$ is the rejection region, one can write:

$$H_0 : (a, b) \in U \quad vs. \quad H_1 : (a, b) \in M \;\; \alpha = \mu, \mathcal{G} = A_1$$

or, symmetrically:

$$H_0 : (a, b) \in M \quad vs. \quad H_1 : (a, b) \in U \;\; \alpha = \lambda, \mathcal{G} = A_2$$

In this formalization, the linkage rule is indeed a likelihood ratio test (Neyman-Pearson) which is the uniformly most powerful test that can be designed for a couple of hypotheses.

At this point it is even more clear that fixing the values of μ and λ (i.e., the significance) is equivalent to bound the admissible error in the decision process.

Searching methods	Comparison functions	Decision models
Sorted Neighborhood	Hamming distance	Probabilistic
Blocking	Edit distance	– EM based
– Sorting	Jaro's algorithm	– Error based
– Hashing	N-grams	– Cost based
		Induction
		Clustering
		Hybrid

Table 1: Taxonomic dimensions to classify record linkage algorithms.

2.2 Classifying Record Linkage Techniques

Even though there are different approaches to the record linkage problem, a standard layout for a generic algorithm can be detailed:

probability estimation — the probabilities $m(\gamma)$ and $u(\gamma)$ are estimated,

weight computation — the estimations computed at the previous step are used to calculate the weight associated to the comparison vector γ,

weight aggregation — a composite score is computed using an aggregation function which takes all the weights in input,

decision — each record pair is classified into either M or U, according to the value of the composite score and the threshold levels chosen.

This basic layout underlines the essential components that are relevant to compare and classify different probabilistic record linkage methods; this is by no means complete nor exhaustive. Even though it would be out from the scope of this paper, Table [Elfeky et al., 2002] reports a wider range of taxonomic dimensions.

In Section 3 we will refer to the aforementioned dimensions to present the selected approaches; more precisely we will investigates two different Bayesian decision models, error based and cost based, two different weight estimation algorithms, Jaro and Winkler, and a brief reviews of other models that are not derived from the FS theory.

3 Existing Models and Techniques

In this section we present review of the selected approaches based on the probabilistic framework introduced. Firstly, we distinguish between the two main Bayesian decision models, error based and cost based; then we investigate two different approaches to calculate the FS weights (see Section 2). Moreover, other methods along a slightly different direction are reviewed as well.

3.1 Bayesian Decision Models

Both the FS theory and [Newcombe and Kennedy, 1962] are based on the Bayes theorem to calculate suitable probabilities used to decide whether or not two records refer to the same entity according to user-set thresholds. The hypothesis under which the following decision models can be applied is that the conditional *Probability Density Function* (PDF) and the *a priori* matching probabilities must be known.

3.1.1 Error Based

This model is *error based* since it calculates the decision thresholds T_μ and T_λ by minimizing the error of incorrectly classify a record in either M or U. The record pairs in $A \times B$ are sorted according to their composite weights and indexed according to such order. Instead of referring to the record pairs, one can refer to the elements of comparison vector $\gamma \in \Gamma$ without losing generality.

The method ensures that the level of user-defined errors (μ, λ) are admissible. Given the above defined ordering, two indexes are chosen n and n' such that the action taken among A_1, A_3 ensures the minimum error; if no better decision is achievable then A_2 is chosen. This intuition is formalized in the following condition for choosing n and n':

$$\sum_{i=n'}^{|\Gamma|} m_i \geq \lambda > \sum_{i=n+1}^{|\Gamma|} m_i$$

$$\sum_{i=1}^{n-1} u_i < \mu \leq \sum_{i=1}^{n} u_i$$

Under the assumption of mutual statistical independency of the components of γ w.r.t. each of the conditional distributions, the above conditions defines the decision function for each γ_i:

$$d(\gamma_i) = \begin{cases} (1,0,0) & 1 \leq i \leq n \\ (0,1,0) & n < i < n' \\ (0,0,1) & n' \leq i \leq |\Gamma| \end{cases}$$

In practice, when the data sets do not represent real random samples of the whole population, $m(\gamma)$ and $u(\gamma)$ (and thus, weights) can be calculated by two different methods proposed in [Fellegi and Sunter, 1969]. Furthermore, the authors also recall that such probability values can be used on subsequent linkage processes working on subpopulations $A' \subseteq A$, $B' \subseteq B$ if the underlying process which generates the records is the same (i.e., if the data sets are drawn from the same source of information).

Using prior information This method assumes that *a priori* information is available. In other words, it assumes that the probability distributions of both (i) the errors contained in the original records and (ii) the comparison characteristics are known for A and B.

Indeed, the method estimates the respective error-free frequencies of each record field in A and B, denoted as $f_{(\cdot)}$.

For instance, if the field under consideration is "address", it is required to count all the records in which such field is reported correctly. Using the respective counts $N_A, N_B, N_{A\cap B}$, the frequencies of each distinct address are estimated $f_{A_1}, f_{A_2}, \ldots, f_{A_m}$, $f_{B_1}, f_{B_2}, \ldots, f_m,\ f_{(A\cap B)_1}, f_{(A\cap B)_2}, \ldots, f_{(A\cap B)_m}$. Each A_k correspond to a record in A where the field "address" is identified by "1", while similarly A_k correspond to a record where the field "address" is identified by "1". The same holds for each $(A \cap B)_k$ but the counts span to the intersection of the two record sets.

Given the above frequencies and total counts, the authors provide examples on how to estimate the probabilities $m(\cdot)$ and $u(\cdot)$; however, the following *a priori* probabilities of error are required:

- $\varepsilon_A, \varepsilon_B$ probability of misreporting an address into either the two files, A or B;
- $\varepsilon_{A^-}, \varepsilon_{B^-}$ probability of not reporting an address into either the two files, A or B;
- ε_{AB} probability of reporting the address in the wrong set, regardless of the correctness of the value itself.

This point is critical in our opinion. [Fellegi and Sunter, 1969] assume that all the addresses (i.e., different values for a field) have the same probability of being reported erroneously. However, it is not uncommon that complicated addresses are more likely to be mistyped/misreported; in addition, there are many factor, more or less related to the data itself, influencing the probability of error which is all but uniform among different values. Considering the following comparison vector:

$$\gamma = [\text{"addresses disagree"},\ \text{"addresses missing on either file"}]$$

the actual probabilities are composed by means of the above defined error rates:

$$\begin{aligned}
m(\gamma^1) &= [1-(1-\varepsilon_A)(1-\varepsilon_B)(1-\varepsilon_{AB})](\varepsilon_{A^-})(1-\varepsilon_{B^-}) = \\
&= \varepsilon_A + \varepsilon_B + \varepsilon_{AB} \\
m(\gamma^2) &= 1-(1-\varepsilon_{A^-})(1-\varepsilon_{B^-}) = \\
&= \varepsilon_{A^-} + \varepsilon_{B^-} \\
u(\gamma^1) &= \left[1-(1-\varepsilon_A)(1-\varepsilon_B)(1-\varepsilon_{AB})\sum_j \frac{f_{A_j}}{N_A}\frac{f_{B_j}}{N_B}\right](1-\varepsilon_{A^-})(1-\varepsilon_{B^-}) = \\
u(\gamma^2) &= 1-(\varepsilon_{A^-})(1-\varepsilon_{B^-}) = \\
&= \varepsilon_{A^-} + \varepsilon_{B^-}
\end{aligned}$$

If two files are large enough and they are drawn from the same population one may assume that $\frac{f_{A_j}}{N_A} = \frac{f_{B_j}}{N_B} = \frac{f_{(A\cap B)_j}}{N_{AB}}$. From a quantitative point of view, positive weighs

contribute to a "match" decision while negative ones contribute to the opposite. Also, it must be noticed that the weight of each field represents somehow the "rarity" of a value: the rarer the value, the larger the weight. Finally, missing values tend to zero out the weights.

To make this algorithm feasible in practice the authors point out that it is not required to list all possible values for each field but, for the reason outlined above, the portion of the most common one will lead to an optimal approximation.

Probability estimation This method estimates $m(\gamma)$ and $u(\gamma)$ from the available records and it is presented in [Fellegi and Sunter, 1969]. Not only it requires the independence assumption but the two data sets must be large enough to make the estimates valid and statistically significant. Beside the estimates of the probabilities, the algorithm also outputs the number N of linked records.

The algorithm proposed by the author is direct and can be applied simply by instantiating the given formulae with certain frequencies parameters which can be automatically calculated from data. The assumption is that m and u must be such that:

$$\begin{aligned} m(\gamma) &= m_1(\gamma^1) \cdot m_2(\gamma^2) \cdots m_k(\gamma^K);\ K \geq 3 \\ u(\gamma) &= u_1(\gamma^1) \cdot u_2(\gamma^2) \cdots u_k(\gamma^K);\ K \geq 3 \end{aligned}$$

which means that γ must have at least three components and they have to be independent to each other. The frequencies of each different configuration of γ is calculated by direct comparison of A against B; the only frequencies of interest are those of "agreements" configurations, denoted with $\Gamma_h^+ \in \Gamma$ for the hth component. More precisely:

- $\hat{F}_{M_{h^-}}$ = frequency of agreements in all components except the hth and any configuration in the hth component. The associated probability is denoted as $m_h = \sum_{\gamma \in \Gamma_h^+} m(\gamma)$.
- $\hat{F}_{U_h}$ = frequency of agreements in the hth component and any configuration in all but the hth. The associated probability is denoted as $u_h = \sum_{\gamma \in \Gamma_h^+} u(\gamma)$.
- $\hat{F}_M$ = frequency of agreements in all components;

The authors have proven that given the frequencies expressed in terms of m and u

$$\begin{aligned} \hat{F}_{M_{h^-}} &= \frac{N}{N_A N_B} \prod_{j=1, j\neq h}^{3} m_j + \frac{N_A N_B - N}{N_A N_B} \prod_{j=1, j\neq h}^{3} u_j \quad h = 1, 2, 3 \\ \hat{F}_{U_h} &= \frac{N}{N_A N_B} m_h + \frac{N_A N_B - N}{N_A N_B} u_h \quad h = 1, 2, 3 \\ \hat{F}_M &= \frac{N}{N_A N_B} \prod_{j=1}^{3} m_j + \frac{N_A N_B - N}{N_A N_B} \prod_{j=1}^{3} u_j \end{aligned}$$

	Error based	
	Prior Information (PI)	**Probability Estimation (PE)**
Hypotheses	Known PDF	$\|\gamma\| \geq 3$ (independent)
Input	error probabilities	frequency of values
Output	$m(\cdot), u(\cdot), N$	$m(\cdot), u(\cdot)$

Table 2: Brief comparison of differences, similarities and peculiarities of the two methods for weights calculation in error based record linkage techniques. Note: N here indicates the number of matching records.

and solving such equations in m_h, u_h and N, the estimates of $m(\gamma^k)$ and $u(\gamma^k)$ can be computed after the direct observation of $\hat{F}_{M_{h^-}}$, $\hat{F}_{U_1}$, $\hat{F}_{U_2}$ and $\hat{F}_{U_3}$ for the specific configurations γ_i^k, γ_i^1, γ_i^2 and γ_i^3, respectively. As the authors stress, this method requires the sample to be large and representative of the whole population.

Comparison In this paragraph we compare the two techniques that have been reported in Section 3.1.1 and Section 3.1.1, summarized in Table 2.

3.1.2 Cost Based

The cost based model can be seen as a generalization of the classical purely Bayesian decision model we mentioned in the previous section. Instead of relating a link between records only with a probability, this models attaches a *cost function* to each decision (i.e., match vs. non-match). Thus, instead of minimizing the error, this method give hints in designing decision rules based on the minimization of a cost.

Generally speaking, the "cost" models the fact that a misclassification has different impacts on the organization data, depending on many factors influencing the whole data flow.

Linear loss method [Tepping, 1968] proposed a method based on a linear loss function, $g(A_i, (a, b))$, defined for each action A_i on the pair (a, b). Given the conditional probability $P(M|\gamma) = P((a, b) \in M|\gamma[\alpha(a), \beta(b)])$, defined as above, the authors define the expected loss G as a function of the action and the conditional probability: $G(A_i, P(M|\gamma)) = \mathrm{E}[g(A_i, (a, b))]$. Hence, the total expected loss is $\sum P(\gamma) \cdot G(A_i, P(M|\gamma))$, which is minimized in order to obtain the optimal linkage rule.

The authors have shown that under the assumption of linearity of G, the interval $(0, 1)$ for the probability of a match is partitioned into a fixed number of possible actions (e.g., $4, A_1, A_2, A_3, A_4$ but it could be any number). The so called *action interval* is the interval in which the loss function G is minimal w.r.t. the same function evaluated in all other action:

$$G(P) = \min_{A_i} G(A_i, P(M|\gamma))$$

so, in the case of three actions, the decision rule is similar to $d(\cdot)$ presented in Section 3.1.1. Here is an example taken from [Tepping, 1968] (Fig. 1):

$$\begin{aligned}
&\text{Take action } A_4 \text{ if } 0 \leq P(M|\gamma) \leq P_1 \\
&\text{Take action } A_2 \text{ if } P_1 < P(M|\gamma) \leq P_2 \\
&\text{Take action } A_4 \text{ if } P_2 < P(M|\gamma) \leq 1
\end{aligned}$$

Even if the authors do not show the modifications, this method can be adapted without the hypothesis of linearity of the loss function; however, the above intervals would not be intervals in general.

Cost matrix method The approach presented in [Verykios et al., 2003] can be interpreted as a generalization of the early effort proposed by [Tepping, 1968]. Misclassification costs are stored into a so called *cost matrix* C; a single element of C is $c_{i,j}$ where $i \in \{1, 2, 3\}$ (i.e., actions A_1, A_2, A_3, respectively) is the predicted class while $j \in \{0, 1\}$ (i.e., classes M, U, respectively) is the actual class the sample belongs to. Even though it can be done automatically, populating the cost matrix is application specific and often requires domain experts.

In this model, each cost value is twofold. The first component of the cost has to do with the decision process itself (e.g., what is the cost of collecting the information required to undertake a certain decision?); the second part is related to the cost of the consequences of a decision. Under the hypothesis of knowing the PDF f_j for each jth component of the comparison vector, this method minimizes the average cost $\bar{c}$ for a given action. If the cost matrix is:

$$C = \begin{pmatrix} c_{A_1M} & c_{A_1U} \\ c_{A_2M} & c_{A_2U} \\ c_{A_3M} & c_{A_3U} \end{pmatrix}$$

the average total cost results in:

$$\begin{aligned}
\bar{c} &= c_{A_1M}P(A_1|M) + c_{A_1U}P(A_1|U) + \\
&+ c_{A_2M}P(A_2|M) + c_{A_2U}P(A_2|U) + \\
&+ c_{A_3M}P(A_3|M) + c_{A_3U}P(A_3|U).
\end{aligned}$$

Knowing that $\pi_M = P(M)$ and $\pi_U = P(U)$ are the *a priori* probabilities of matching, then:

$$\begin{aligned}
\bar{c} &= c_{A_1M}P(A_1|M)\pi_M + c_{A_1U}P(A_1|U)\pi_U + \\
&+ c_{A_2M}P(A_2|M)\pi_M + c_{A_2U}P(A_2|U)\pi_U + \\
&+ c_{A_3M}P(A_3|M)\pi_M + c_{A_3U}P(A_3|U)\pi_U.
\end{aligned}$$

	Cost based	
	Linear Loss Function (LLF)	**Cost Matrix (CM)**
Hypotheses	Loss as a linear function	Known PDF
Input	Conditional probabilities	Per-action cost
Output	Decision intervals	Decision thresholds

Table 3: Brief comparison of differences, similarities and peculiarities of the two methods for cost based record linkage techniques. Note: N here indicates the number of matching records.

Skipping the details reported in [Tepping, 1968], it can be proved that three thresholds exist in the decision space. In particular, referring to λ, μ defined by the FS theory the thresholds are:

$$\lambda = \frac{\pi_U}{\pi_M} \cdot \frac{c_{A_2M} - c_{A_1M}}{c_{A_1U} - c_{A_2U}} \quad \kappa = \frac{\pi_U}{\pi_M} \cdot \frac{c_{A_3M} - c_{A_1M}}{c_{A_1U} - c_{A_3U}} \quad \mu = \frac{\pi_U}{\pi_M} \cdot \frac{c_{A_3M} - c_{A_2M}}{c_{A_2U} - c_{A_3U}}.$$

According to the actual values of such thresholds, the decision space depicted in Figure 1 is divided into two or three areas. In particular, the authors show that the sufficient and necessary condition for A_2 to exist is that $\lambda \leq \mu$. Furthermore, if it holds, they have shown that $\lambda \leq \kappa \leq \mu$. Otherwise (i.e., $\lambda > \mu$) $\lambda > \kappa > \mu$ but κ is such that A_2 disappears and there are only two decision areas. This happens because A_2 have a higher cost w.r.t. A_1 and A_3 because it results in manual classification.

However, the method is proven by the authors to be optimal w.r.t. the cost, in the sense that it minimizes a cost function; but on the other hand, no proof is given of its optimality in general. In other words, no clues are given to prefer this method over the others available.

Comparison In this paragraph we compare the two techniques that have been reported in Section 3.1.2 and Section 3.1.2, summarized in Table 3.

The method by [Tepping, 1968] requires slightly stronger assumption w.r.t. to the one recently proposed in [Verykios et al., 2003]. However, the idea of using cost as a criterion was originally due to [Tepping, 1968], in which the theoretical framework has been defined and detailed. The rigorous theory proposed in [Fellegi and Sunter, 1969] was significant in the construction of the more complete and directly applicable approach by [Verykios et al., 2003].

3.1.3 Comparing Error Based and Cost Based Decision Models

It is interesting to compare the two decision models at a generic level. In particular, it could be proven that the former is a special case of the latter. First of all, it must be remarked that they are both likelihood ratio tests with thresholds computed on the basis of the available information, that is the *a priori* probabilities.

	Error based		Cost based	
	PI	**PE**	**LLF**	**CM**
Hypotheses	Known PDF	$\|\gamma\| \geq 3$ (independent)	Loss as a linear function	Known PDF
Input	error probabilities	frequency of values	Conditional probabilities	Per-action cost
Output	$m(\cdot), u(\cdot), N$	$m(\cdot), u(\cdot)$	Decision intervals	Decision thresholds

Table 4: Comparison of differences, similarities and peculiarities of error based vs. cost based record linking methods.

As we pointed out in Section 2, in the case of error minimization the likelihood test can be computation of thresholds relies on the following ratio:

$$l(\gamma) = \frac{m(\gamma)}{u(\gamma)} \quad T_\mu = l(\gamma_n), T_\lambda = l(\gamma_{n'})$$

while on the other hand, the cost minimization criterion can be reduced to the estimation of

$$l^c(\gamma) = \frac{(c_{21} - c_{11}) \cdot m^c(\gamma)}{(c_{12} - c_{22}) \cdot u^c(\gamma)} \quad T^c_\mu = l^c(\gamma_n), T^c_\lambda = l^c(\gamma_{n'})$$

If the cost of a misclassification is equal $c_{12} - c_{22} = c_{21} - c_{11}$ in both the cases then $l = l^c$. In other words, the selection of the cost function is equivalent to changing the *a priori* probabilities.

Finally, it curious to notice how the two methods have been compared together [Tepping, 1968]: the authors claim that cost based methods are better than error based methods, when no further clue is available to properly assign error probabilities. Contradictorily, in the experimental results, they show and conclude that the error probability is *always lower* if error based methods are used instead of cost based methods.

Table 4 summarizes the differences, the similarities and the peculiarities of the two approaches.

3.1.4 EM Algorithm Based

In this section, a technique using the *Expectation Maximization* (EM) algorithm [Dempster et al., 1977] is investigated. We separate the review of the method based on this algorithm from the others because it is along a different line and it has been developed using a generic and widely-applied estimation method. However, in Section 3.1.5, a comprehensive comparison is presented including this technique.

The use of the EM algorithm for record linkage have been proposed by [Jaro, 1989] and recently re-investigated in [Winkler, 1988]. It is based on likelihood estimators and it can be potentially used in any kind of probabilistic model to find the parameters, even

in the presence of unobservable variables or missing data. Given these premises, it is straightforward to notice that the EM algorithms perfectly fits the probabilistic model defined by the FS theory.

To avoid misunderstandings, we will use the $\underline{v}$ to indicate that v is a vector. The parameters of interests are $\Phi = \langle \underline{m}, \underline{u}, p \rangle$ where $\underline{m}, \underline{u}$ denotes the probability vectors $m(\cdot), u(\cdot)$, respectively; and, p denotes the proportion of the matched records w.r.t. the total: $p = \frac{|M|}{|M \cup U|}$. The data vector is defined by γ and the function g:

$$g_j = \begin{cases} (1,0) & (a,b)_j \in M \\ (0,1) & (a,b)_j \in U \end{cases}$$

The data vector is then $\underline{x} = \langle \gamma, g \rangle$; we recall that $(a,b)_j$ indicates the generic jth record pair. An independence model is assumed at this step, thus:

$$\begin{aligned} P(\gamma^j|M) &= \prod_{i=1}^{n} m_i(\gamma_i^j)(1-m_i)^{1-\gamma_i^j} \\ P(\gamma^j|U) &= \prod_{i=1}^{n} u_i(\gamma_i^j)(1-u_i)^{1-\gamma_i^j} \end{aligned}$$

Given the log-likelihood of the data vector:

$$\begin{aligned} \log f(\underline{x}|\Phi) &= \sum_{j=1}^{N} g_j \cdot (\log P(\gamma^j|M), \log P(\gamma|U))^T + \\ &+ \sum_{j=1}^{N} g_j \cdot (\log p, \log(1-p))^T \end{aligned}$$

the algorithm consists in the iteration of two steps called *Expectation* (E) and *Maximization* (M); the iteration begins with the initial (even casual) estimates $\langle \hat{\underline{m}}, \hat{\underline{u}}, \hat{p} \rangle$ continues until the required precision is not reached. The estimation of $\underline{u}$ is less difficult w.r.t. $\underline{u}$ since $|U| > |M|$, thus the u_i can be estimated by ignoring the contribution of M. Regarding m, the g function can be estimated in the (E) step as follows:

$$\begin{aligned} \hat{g}_m(\gamma^j) &= \frac{\hat{p} \cdot P(\gamma^j|M)}{\hat{p} \cdot P(\gamma^j|M) + (1-\hat{p}) \cdot P(\gamma^j|U)} \\ \hat{g}_u(\gamma^j) &= \frac{\hat{p} \cdot P(\gamma^j|U)}{\hat{p} \cdot P(\gamma^j|U) + (1-\hat{p}) \cdot P(\gamma^j|M)} \end{aligned}$$

Note that g_j is estimated by $\langle \hat{g}_m(\gamma^j), \hat{g}_m(\gamma^j) \rangle$. The (M) step, in the case of $\hat{\underline{m}}$, it is based on:

$$\hat{m}_j = \frac{\sum_{j=1}^{s^n} \hat{g}_m(\gamma^j) \gamma_i^j \hat{F}(\gamma^j)}{\sum_{j=1}^{s^n} \hat{g}_m(\gamma^j) \hat{F}(\gamma^j)}$$

	Error based		Cost based		EM Based
	PI	**PE**	**LLF**	**CM**	
Hypotheses	Known PDF	$\|\gamma\| \geq 3$ (independent)	Loss as a linear function	Known PDF	Independency
Input	error probabilities	frequency of values	Conditional probabilities	Per-action cost	frequency of values and cond. prob.
Output	$m(\cdot), u(\cdot), N$	$m(\cdot), u(\cdot)$	Decision intervals	Decision thresholds	$\hat{m}(\cdot), \hat{u}(\cdot), \hat{p}$

Table 5: Summary of the main methods for parameter estimation used in probabilistic record linkage techniques.

where $\hat{F}(\gamma^j)$ indicates the frequency count for the jth component of the comparison vector γ. The estimate of p is then $\hat{p} = \frac{\sum_{j=1}^{s^n} g_m(\gamma^j)\hat{F}(\gamma^j)}{\sum_{j=1}^{s^n} \hat{F}(\gamma^j)}$.

The author underlines that the method is extremely easy to implement, stable, and negligibly sensitive to initialization values. It is also highlighted that all the frequency counts have to be obtained after a blocking phase, but, this detail is out from the scope of this survey so we refer the reader to [Cochinwala et al., 2001] for a more general overview of all the steps of a liking algorithm.

This algorithm have been used to detect duplicates in the census data of Tampa, Florida in 1985 [Jaro, 1989] and lately on public health data [Jaro, 1995].

3.1.5 Overall comparison

Cost based methods have been already compared with error based techniques. Table 5 summarizes all the (variants of the) approaches that have been reviewed.

One may have noticed that the EM method shares slightly the same hypotheses and the same input/output w.r.t. the PE method. However, the latter also requires the comparison vector γ to have at least three components while the former does not have this limitation.

3.2 Other Models and Methods

In the previous sections we investigated and reviewed the most solid and promising methods that are also implemented into bleeding edge tools (see Section 4). For the sake of completeness, in this section we provide list of other approaches that have been proposed in the literature so far.

In particular, a different way to model the error in the data is to *explicitly* model the errors in *each* attribute, as proposed in [Copas and Hilton, 1990]. The algorithm is based on the statistical characteristics of the errors that are expected to arise; however,

the error distribution needs large training data sets of *already matched* pairs in order to be correctly estimated.

Also, in [Larsen and Rubin, 2001] a Bayesian method is proposed based on mixture models and marginal information; the method allows to label manually reviewed data, if available.

In [Fortini et al., 2001] another Bayesian model is proposed using two different algorithms to derive the marginal posterior distribution of the configuration of matches between the two data sets.

Along a different line, the methods reported in [Bilenko and Mooney, 2002, Pinheiro and Sun, 1998, Tejada et al., 2001] uses a prediction approach to estimate model parameters. In particular, [Bilenko and Mooney, 2002] is based on support vector machines, [Pinheiro and Sun, 1998] relies on logistic regression and [Tejada et al., 2001] utilizes decision trees. As pointed out by [Winkler, 1999], such algorithms requires a non-negligible amount of training data.

Finally, Du Bois Jr [1969] shows how the precision of a probabilistic matching algorithm can be increased if *two* comparison vectors are used. In particular, instead of taking into account the agreement vector only, the author propose to also build a *presence* vector and to aggregate them both into a more detailed comparison indicator. More precisely, two variables X_j and Y_j are defined for each records pair: $X_j = 1$ only if the jth corresponding item on both records is *present*, zero otherwise. $Y_j = 1$ only if the jth corresponding item on both records agrees, zero otherwise. Then the composite random variable X_jY_j is defined: as the intuition suggests, $X_jY_j = 1$ only if both $X_j = 1$ and $Y_j = 1$, zero otherwise. Without going into the details, it possible to estimate the model parameters if each component of the composite vector is supposed to be binomially distributed on M and U.

4 Existing tools

In this section we provide a short taxonomic list of the available data linkage, scrubbing, and cleansing software. Both commercial products and research prototypes are included. This list is not meant to be complete nor exhaustive: it s rather an overview of the most recent and promising tools focused on probabilistic algorithms. Also, we selected two applications for which a more detailed analysis is presented, namely TAILOR [Elfeky et al., 2002] and Febrl [Christen, 2008]. We also refer the reader to [Gu et al., 2003] for another (tool) survey.

The formerly known as Integrity developed by Vality is now maintained by IBM into the WebSphere project. It is now named Integrator [IBM Corporation, 2005] and includes lexical analysis, pattern processing, statistical matching and data streaming. WebSphere Integrator is customizable w.r.t. the business rules; it also includes a statistical algorithm for record matching.

The Trillium software [Harte-Hanks] is a four modules tool with good features for data analysis. The matching module is particularly precise and accurate in linking records containing addresses, thanks to a geocoder module.

The LinkageWiz software [Data Quality Solutions] has been developed with the specific purpose of assessing the data quality of health records. It has been then released under a commercial license. Beside phonetic name matching, nickname to name mapping, and data quality indicators, LinkageWiz provides the user to tune the probabilistic matching algorithms by specifying the matching weights directly.

The commercial database reconciliation tool developed by Telcordia.com [Caruso et al., 2000] allows the generation of custom rules, subsequently run on the whole database to assess the data quality and/or to identify matching records. It has been used to detect duplicates records in the database of taxpayers.

The *Generalized Record Linkage System* (GRLS) [Fair, 2001] includes a full implementation of the Fellegi-Sunter theory within a graphical interface. A useful feature of this tool is that it allows the user to tune the matching algorithm with given thresholds and weights; also, it is also possible to plug custom rules into the matching engine. Finally, differently from other systems, it permits to group matched records into *two* groups according to the actual scoring: weak matches and strong matches.

The WizRule [WizSoft] software by WizSoft is not a record linker tool but the results of its analysis can be further processed to identify linked records. Indeed, it is a rule mining and discovery application which exhaustively search the records for association rules.

4.1 Febrl: A Comprehensive, Open Source Platform for Record Linkage

Febrl [Christen, 2008], or Freely Extensible Biomedical Record Linkage, implements the state-of-the art algorithms for record linkage (`http://febrl.sf.net`). Febrl offers the researchers a fully pluggable programming interface written in Python (one of the most easy-to-learn programming languages) encapsulated into a full fledged graphical interface. The relevance of this tool for the scientific community is due to its extensibility and flexibility w.r.t. new algorithms that can be quickly implemented in order to compare their performances against the existing ones.

Not only Febrl provides an open-source testbed for new methods but it ships with a rich collection of classic and recent record linkage algorithms; it also includes some sample data sets and a data set generator.

Beside common functionalities like the support for multi-format input/output, summary for data exploration, and logging, Febrl includes a data cleansing and name/address standardization module which implements a rule-based algorithm in conjunction with an hidden Markov model engine. As for direct field comparison, the tool allows to select among 26 alternative functions including approximate string comparators and ad-hoc procedures for special fields (e.g., addresses, dates). The available decision algorithms are the classic Fellegi & Sunter (with a supervised variant to help the user in the threshold-setting phase), support vector machine, k-means and farthest-first clustering methods, and an unsupervised classification algorithm.

	Febrl	**TAILOR**
Searching	Blocking	Blocking
	– Sorting	– Sorting
	– Suffix array	– Hashing
	– Fuzzy (Q-gram)	
	Sorted neighborhood	Sorted neighborhood
	Full index	
Comparison	Hamming	Hamming
	Edit	Edit
	Winkler	Jaro's
	Q-grams	N-grams
	Soundex	Soundex
	Key-diff	
	Approximate match	
	ad-hoc	
Decision	Probabilistic	Probabilistic
	– EM-based	– EM-based
	– error-based	– cost-based
	– optimal threshold	– error-based
	Support Vector Machine	Induction
	Clustering	Clustering
	Binary classification	Hybrid

Table 6: Feature comparison of Febrl vs. TAILOR.

4.2 TAILOR: A Record Linkage Toolbox

TAILOR, proposed in [Elfeky et al., 2002], is a record linkage toolbox intended to provide the users an extensible and abstract framework to both develop and test record matching algorithms. In such sense, TAILOR is similar to Febrl but it adopts a different approach. Febrl standardizes the programming interface while TAILOR allows to extend its interface to adhere to the one of the existing tool that needs to be integrated. TAILOR also includes a data synthesizing module based on DBGen, a publicly available tool for parametric data generation.

The alternative algorithms for each cleansing phase implemented in TAILOR are summarized in Table 1. As for the decision models, TAILOR implements several algorithms: EM-based, cost-based, error-based, clustering, induction and hybrid. The clustering of records is based on k-means as in Febrl. In addition, TAILOR implements the induction linkage model: it includes both the aforementioned approach based on decision-tree (Section 3.2) and an instance based learning algorithm. The hybrid model refers to the option of using induction together with clustering: the former is supervised and more accurate but it requires labeled training sets which may not be available; the latter is unsupervised and it is exploited to predict the class (label) of each pattern to be analyzed

by the supervised phase.

5 Conclusions

In this paper we have presented a survey of the most recent and promising probabilistic record linkage methods, along with a brief overview of the tools that can be used to accomplish duplicate detection or data cleansing, generally.

Our study highlights that the stochastic approach is promising and that the abstract, rigorous theory proposed by [Fellegi and Sunter, 1969] is applicable in practice even if a few simplifying assumption are still needed. On the other hand, some of the existing methods to estimate the models parameters rely on hypotheses that do not always hold. However, more pragmatic techniques such as the cost based ones allow the user to fine tune the resulting decision according to the expected maximum costs; this can be of help when no other alternative can be applied, for instance because there is no sufficient *a priori* knowledge regarding the data/error generation process.

The widespread of the applications of record linkage techniques is the strongest confirmation of the actual effectiveness of the existing approaches. Indeed, there is a large amount of both free and commercial software available to the institutions: this is another fact confirming that the probabilistic record linkage techniques work. They are not only proof-of-concept prototypes used by the researchers but they are ready for the real world. However, it must be underlined that many of the real applications has to do with typographical variations and slight errors where classical probabilistic techniques perform well on.

Even though we did not addressed the topic of the performance of record matching, according to the reviewed literature speed is still an issue. Linking algorithms needs to be *fast* for practical applications, since (1) in the worst case they work on the *Cartesian product* of the two data set to be compared and (2) data sources under are usually *large* (e.g., census data).

The evaluation of the accuracy of a linkage algorithm is still an open research question [Winkler, 1999] even though some metrics to measure the linking error rates have been proposed; however, the issue is how to suitably automatize them. Other relevant and advanced research problems are reported in [Winkler, 1999].

Finally, side problems that this paper has not discussed are related to ethical issues (e.g., privacy), confidentiality and legal consequences of inaccuracies due to an automatic linking system. For such topics we refer the reader to the references listed in Section 4 of [Gu et al., 2003].

Disclaimer This study has been written as a final exam of the *"Advanced Topics in Information Systems B"* PhD course taught by Prof. B. Pernici at Politecnico di Milano. It is our care to remind that this is *not* a *peer reviewed* work thus it may contain inaccuracies.

References

Mikhail Bilenko and Raymond J. Mooney. Learning to combine trained distance metrics for duplicate detection in databases. Technical report, University of Texas at Austin, 2002.

F. Caruso, M. Cochinwala, U. Ganapathy, G. Lalk, and P. Missier. Telcordia's Database Reconciliation and Data Quality Analysis Tool. *Proceedings of the 26th International Conference on Very Large Data Bases*, pages 615–618, 2000.

P. Christen. Febrl–An Open Source Data Cleaning, Deduplication and Record Linkage System with a Graphical User Interface. August 2008.

M. Cochinwala, S. Dalal, A. K. Elmagarmid, and V. S. Verykios. Record matching: Past, present and future. Technical report, PR-OWL: A Bayesian Ontology Language for the Semantic Web. Workshop on Uncertainty Reasoning for the Semantic Web, International Semantic Web Conference, 2001.

JB Copas and FJ Hilton. Record linkage: statistical models for matching computer records. *Journal of the Royal Statistical Society Series A*, 153:287–320, 1990.

Data Quality Solutions. LinkageWiz. Available online at `http://www.linkagewiz.com`.

A.P. Dempster, N.M. Laird, D.B. Rubin, et al. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

Du Bois Jr. A Solution to the Problem of Linking Multivariate Documents. *Journal of the American Statistical Association*, 64(325):163–174, 1969.

M.G. Elfeky, V.S. Verykios, and A.K. Elmagarmid. Tailor: a record linkage toolbox. *Data Engineering, 2002. Proceedings. 18th International Conference on*, pages 17–28, 2002. doi: 10.1109/ICDE.2002.994694.

M.E. Fair. Recent Developments at Statistics Canada in the linking of complex health files. *Federal Committee on Statistical Methodology, Washington, DC*, 2001.

I.P. Fellegi and A.B. Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210, 1969.

M. Fortini, B. Liseo, A. Nuccitelli, and M. Scanu. On Bayesian record linkage. *Research in Official Statistics*, 4:185–198, 2001.

L. Gu, R. Baxter, D. Vickers, and C. Rainsford. Record linkage: Current practice and future directions. *CSIRO Mathematical and Information Sciences Technical Report*, 3:83, 2003.

Harte-Hanks. Trillium Software System. Available online at `http://www.trilliumsoftware.com/home/products/index.aspx`.

IBM Corporation. IBM WebSphere Information Integrator Version 8.2. Technical report, IBM Corporation, 2005.

M.A. Jaro. Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association*, 84(406), 1989.

MA Jaro. Probabilistic linkage of large public health data files. *Stat Med*, 14(5-7):491–8, 1995.

M.D. Larsen and D.B. Rubin. Iterative Automated Record Linkage Using Mixture Models. *Journal of the American Statistical Association*, 96(453), 2001.

Howard B. Newcombe and James M. Kennedy. Record linkage: making maximum use of the discriminating power of identifying information. *Commun. ACM*, 5(11):563–566, 1962. ISSN 0001-0782. doi: http://doi.acm.org/10.1145/368996.369026.

J.C. Pinheiro and D.X. Sun. Methods for Linking and Mining Massive Heterogeneous Databases. *Knowledge Discovery and Data Mining*, pages 309–313, 1998.

S. Tejada, C.A. Knoblock, and S. Minton. Learning object identification rules for information integration. *Information Systems*, 26(8):607–633, 2001.

B.J. Tepping. A model for optimum linkage of records. *Journal of the American Statistical Association*, 63(324):1321–1332, 1968.

V. S. Verykios, G. V. Moustakides, and M. G. Elfeky. A bayesian decision model for cost optimal record matching. *The VLDB Journal*, 12(1):28–40, 2003. ISSN 1066-8888. doi: http://dx.doi.org/10.1007/s00778-002-0072-y.

W.E. Winkler. Using the EM algorithm for weight computation in the Fellegi-Sunter model of record linkage. *Section on Survey Research Methods*, 1988.

W.E. Winkler. The state of record linkage and current research problems. *Statistics of Income Division, Internal Revenue Service Publication R*, 4, 1999.

WizSoft. WizRule. Available online at `http://www.wizsoft.com/default.asp?win=8&winsub=8`.