A Generalized Cost Optimal Decision Model for Record Matching

Vassilios S. Verykios & George V. Moustakides
Department of Computer and Communication Engineering
University of Thessaly, Volos,
GREECE
verykios@inf.uth.gr moustaki@inf.uth.gr

ABSTRACT

Record (or entity) matching or linkage is the process of identifying records in one or more data sources, that refer to the same real world entity or object. In record linkage, the ultimate goal of a decision model is to provide the decision maker with a tool for making decisions upon the actual matching status of a pair of records (i.e., documents, events, persons, cases, etc.). Existing models of record linkage rely on decision rules that minimize the probability of subjecting a case to clerical review, conditional on the probabilities of erroneous matches and erroneous non-matches. In practice though, (a) the value of an erroneous match is, in many applications, quite different from the value of an erroneous non-match, and (b) the cost and the probability of a misclassification, which is associated with the clerical review, is ignored in this way. In this paper, we present a decision model which is optimal, based on the cost of the record linkage operation, and general enough to accommodate multi-class or multi-decision case studies. We also present an example along with the results from applying the proposed model to large comparison spaces.

Keywords

Record Matching, Probabilistic Decision Model

1. INTRODUCTION

In today's competitive business environment, corporations in the private sector are being driven to focus on their customers in order to maintain and expand their market share. This shift is resulting in customer data and information about customers being viewed as a corporate asset. In the public sector, the very large expansion of the role of the government resulted in an unprecedented increase in the demand for detailed information. Only recently has the data analytical value of these administrative records been fully realized. Of primary concern is that, unlike a purposeful data collection effort, the coding of the data is not carefully controlled for quality. Likewise, data objects are not necessarily defined commonly across databases nor in the way data consumers would want. Two of the serious concerns which arise in this context are (a) how to identify records across different data stores that refer to the same entity and (b) how to identify duplicate records within the same data store.

If each record in a database or a file carried a unique, universal and error-free identification code, the only problem would be to find an optimal search sequence that would minimize the total number of record comparisons. In most cases, encountered in practice, the identification code of the record is neither unique nor error-free. In some of these cases, the evidence presented by the identification codes (i.e., primary key, object id, etc.) may possibly point out that the records correspond or that they do not correspond to the same entity. However, in the large majority of practical problems, the evidence may not clearly point to one or the other of these two decisions. Thus, it becomes necessary to make a decision as to whether or not a given pair of records must be treated as though it corresponds to the same real world entity. This is called the record matching or record linkage problem [6, 1].

The large volume of applications spanning the range of cases from (a) an epidemiologist, who wishes to evaluate the effect of a new cancer treatment by matching information from a collection of medical case studies against a death registry in order to obtain information about the cause and the date of death, to (b) an economist, who wishes to evaluate energy policy decisions by matching a database containing fuel and commodity information for a set of companies against a database containing the values and the types of goods produced by the companies, signifies the tremendous impact and applicability of the problem addressed in this paper.

The remaining of this paper is organized as follows. Section 2 provides some background information, and the notation that is used throughout this paper. Section 3 introduces the cost optimal model, along with the thresholds of the decision areas, and the probabilities of errors. An example is given in Section 4 to illustrate how the model can be applied. Section 5 provides some information about the experimental environment that we generated and the results of some experiments that we run by using it. Finally, Section 6 provides concluding remarks and guidelines for future extensions of this work.

IQIS 2004 Maison de la Chimie, Paris, France
©2004 ACM 1-58113-902-0/04/0006 $5.00.

2. BACKGROUND

Record matching or linking is the process of identifying records, in a data store, that refer to the same real world entity or object. The two principal steps in the record matching process are the searching step where we search for potential linkable pairs of records and the matching step where we decide whether or not a given pair is correctly matched. The aim of the searching step must be to reduce the possibility of failing to bring linkable records together for comparison. For the matching step, the problem is how to enable the computer to decide whether or not a pair of records relates to the same entity, when some of the identifying information agrees and some disagrees.

2.1 Notation

In the product space of two tables, a *match* M is a pair that represents the same entity and a *non-match* U is a pair that represents two different entities. Within a single table, a *duplicate* is a record that represents the same entity as another record in the same database. Common record identifiers such as names, addresses and code numbers (SSN, object identifier), are the matching variables that are used to identify matches. The vector, that keeps the values of all the attribute comparisons for a pair of records (comparison pair) is called *comparison vector* $\underline{x}$. The set of all possible vectors, is called *comparison space* X. A record matching rule is a decision rule that designates a comparison pair either as a *link* A_1, a *possible link* A_2, or a *non-link* A_3, based on the information contained in the comparison vector. Possible links are those pairs for which there is no sufficient identifying information to determine whether a pair is a match, or a non-match. Typically, manual review is required in order to decide upon the matching status of possible links. *False matches* (Type I errors) are those non-matches that are erroneously designated as links by a decision rule. *False non-matches* (Type II errors) are either (a) matches designated as non-links by the decision rule, or (b) matches that are not in the set of pairs to which the decision rule is applied.

For an arbitrary comparison vector $\underline{x} \in X$, we denote by $P(\underline{x} \in X|M)$ or $f_M(\underline{x})$ the frequency of the occurrence or the conditional probability of the particular agreement $\underline{x}$ among the comparison pairs that are matches. Similarly, we denote by $P(\underline{x} \in X|U)$ or $f_U(\underline{x})$ the conditional probability of $\underline{x}$ among the non-matches. Note that the agreement or comparison vector $\underline{x}$ can be defined as specifically as one wishes and this completely rests to the components of the comparison vector. Let p_j be the probability that the j-th corresponding item on the records a and b is present when the outcome of the comparison (a, b) is a match, and let p_j^* be similarly defined when the outcome is a non-match. Likewise, let q_j be the probability that the j-th corresponding item on the records a and b is identical when the outcome of the comparison (a, b) is a match and let q_j^* be similarly defined when the outcome is a true non linkage. Let us also denote by $P(d = A_i, r = j)$ and $P(d = A_i|r = j)$ correspondingly, the joint and the conditional probability that the decision A_i is taken, when the actual matching status (M or U) is j. We also denote by c_{ij} the cost of making a decision A_i when the comparison record corresponds to some pair of records with actual matching status j. When the dependence on the comparison vector is obvious from the context, we eliminate the symbol $\underline{x}$ from the probabilities. Finally we denote the a-priori probability of M or else $P(r = M)$ as π_0 and the a-priori probability of U or else $P(r = U)$ as $1 - \pi_0$.

2.2 Decision Models for Record Matching

In 1950s, Newcombe et. al. [9] introduced concepts of record matching that were formalized in the mathematical model of Fellegi and Sunter [2]. Newcombe recognized that linkage is a statistical problem: in the presence of errors of identifying information to decide which record pair of potential comparisons should be regarded as linked. Fellegi and Sunter formalized this intuitive recognition by defining a linkage rule as a partitioning of the comparison space into the so-called "linked" subset, a second subset for which the inference is that the record pairs refer to different underlying units and a complementary third set where the inference cannot be made without further evidence.

Fellegi and Sunter in [2], making rigorous concepts introduced by Newcombe et. al. [10] considered ratios of probabilities of the form:

$$R = P(\underline{x} \in X|M)/P(\underline{x} \in X|U) \quad (1)$$

where $\underline{x}$ is an arbitrary agreement pattern in the comparison space X. The theoretical decision rule is given by:

```
(a) If $R >$ UPPER, then designate pair as link.
(b) If LOWER $\leq R \leq$ UPPER, then designate the
        pair as a possible link and hold it
        for clerical review.
(c) If $R <$ LOWER, then designate the pair
        as non-link.
```

The UPPER and LOWER cutoff thresholds are determined by a-priori error bounds on false matches and false non-matches. Fellegi and Sunter [2] showed that the decision rule is optimal in the sense that for any pair of fixed upper bounds on the rates of false matches and false non-matches, the manual/clerical review region is minimized over all decision rules on the same comparison space X. If now, one considers the costs of the various actions, that might be taken, and the utilities associated with their possible outcomes, it is desirable to choose decision rules that will minimize the costs of the operation. Tepping in [11] provides a graphical representation of a solution methodology that minimizes the mean value of the cost under the condition that the expected value of the loss is a linear function of the conditional probability that the comparison pair is a match. The application of his mathematical model involves the estimation of the cost function for each action, as a function of the probability of a match, and the estimation of the probability that a comparison pair is a match.

3. THE GENERALIZED COST OPTIMAL DECISION MODEL

Here, we propose a new cost optimal decision model for record matching. The model presented here is a generalization of the model that it was proposed in [13] in the sense that the number of decision areas (link, non-link, possible link) is not restricted to three but it can be any non-negative number n. In general, let us denote by c_i^j the cost of making a decision A_i for a comparison pair in the state of nature

j. Each one of the decisions that are made, based on the existing evidence, about the linking status of a comparison pair, is associated with a certain cost that has two aspects. The first aspect is related to the decision process itself and is associated with the cost of making a particular decision; for example, the number of value comparisons that are needed in order to decide, affects the cost of this decision. The second aspect is associated with the cost of the impact of a certain decision; for example, making a wrong decision should always cost more than making the correct decision. Table 1 illustrates the costs for all the various decisions that could be made during the record matching process.

Table 1: Costs of the decisions.

Cost	Decision	State of Nature
c_1^M	A_1	M
c_1^U	A_1	U
c_2^M	A_2	M
c_2^U	A_2	U
...	...	...
c_n^M	A_n	M
c_n^U	A_n	U

A record linkage process assigns each one of the comparison pairs to one and only one decision area. In order to compute the mean cost of the record linkage process, we consider one by one the costs of all decision areas. Without loss of generality, let us consider the cost of the i-th decision area. What we know about this area is that it has been assigned a number of comparison vectors based on a decision process that we are trying to identify. It is also the case, that among the comparison vectors allocated to this area, there maybe both matched and non-matched comparison pairs. There is a certain probability measure about the fact that a comparison pair (matched or non-matched) is allocated to this decision area. This is denoted by the joint probability $P(d = A_i, r = M)$ and $P(d = A_i, r = U)$ respectively. For every matched comparison pair assigned to the decision area i the associated cost is c_i^M and for every non-matched comparison pair assigned to this area, the cost is c_i^U. The mean cost over all decision areas can then be written as follows:

$$\overline{c} = \sum_{i=1}^{n} [c_i^M \cdot P(d = A_i, r = M) + c_i^U \cdot P(d = A_i, r = U)] \quad (2)$$

We can express the joint probabilities in Eq. 2 as a function of the conditional probabilities by using the Bayes theorem. Based on this observation, for $i = 1, 2, \ldots, n$ and $j = M, U$, we get:

$$P(d = A_i, r = j) = P(d = A_i | r = j) \cdot P(r = j). \quad (3)$$

Let us also assume that $\underline{x}$ is a comparison vector drawn randomly from the space of the comparison vectors which is shown in Figure 1. Then the following equality holds for the conditional probability $P(d = A_i | r = j)$:

$$P(d = A_i | r = j) = \sum_{\underline{x} \in A_i} f_j(\underline{x}), \; i = 1, 2, \cdots, n; \; j = M, U, \quad (4)$$

where f_j is the probability density of the comparison vectors when the state of nature is j. We also denote the a-priori probability of M or else $P(r = M)$ by π_0 and the a-priori probability of U or else $P(r = U)$ as $1 - \pi_0$.

The mean cost $\overline{c}$ in Eq. 2 based on Eq. 3 is written as follows:

$$\overline{c} = \sum_{i=1}^{n} [c_i^M \cdot P(d = A_i | r = M) \cdot P(r = M) + c_i^U \cdot P(d = A_i | r = U) \cdot P(r = U)]. \quad (5)$$

By using Eq. 4, Eq. 5 becomes:

$$\overline{c} = \sum_{i=1}^{n} [c_i^M \cdot \sum_{\underline{x} \in A_i} f_M(\underline{x}) \cdot P(r = M) + c_i^U \cdot \sum_{\underline{x} \in A_i} f_U(\underline{x}) \cdot P(r = U)] \quad (6)$$

By substituting the a-priori probabilities of M and U in Eq. 6, we get the following equation:

$$\overline{c} = \sum_{i=1}^{n} [c_i^M \cdot \pi_0 \cdot \sum_{\underline{x} \in A_i} f_M(\underline{x}) + c_i^U \cdot (1 - \pi_0) \cdot \sum_{\underline{x} \in A_i} f_U(\underline{x})] \quad (7)$$

which by dropping the dependent vector variable $\underline{x}$, and combining the information for each part of the decision space, can be rewritten as follows:

$$\overline{c} = \sum_{i=1}^{n} \sum_{\underline{x} \in A_i} [f_M \cdot c_i^M \cdot \pi_0 + f_U \cdot c_i^U \cdot (1 - \pi_0)] \quad (8)$$

Every point $\underline{x}$ in the decision space A, belongs either in partition A_1, or in A_2, ..., or in A_n and it contributes additively to the mean cost $\overline{c}$. We can thus assign each point independently either to A_1, or A_2, ..., or A_n in such a way that its contribution to the mean cost is minimal. This will lead to the optimum selection for the sets which we denote by A_1^o, A_2^o, ..., and A_n^o. Based on this observation, a point $\underline{x}$ is assigned to the optimal decision area A_i^o iff the following $n - 1$ inequalities hold:

$$\begin{aligned} f_M \cdot c_i^M \cdot \pi_0 + f_U \cdot c_i^U (1 - \pi_0) &\leq f_M \cdot c_1^M \cdot \pi_0 + f_U \cdot c_1^U (1 - \pi_0) \\ f_M \cdot c_i^M \cdot \pi_0 + f_U \cdot c_i^U (1 - \pi_0) &\leq f_M \cdot c_2^M \cdot \pi_0 + f_U \cdot c_2^U (1 - \pi_0) \\ &\vdots \\ f_M \cdot c_i^M \cdot \pi_0 + f_U \cdot c_i^U (1 - \pi_0) &\leq f_M \cdot c_n^M \cdot \pi_0 + f_U \cdot c_n^U (1 - \pi_0) \end{aligned}$$

We thus conclude from the above that for any value of i, the corresponding decision area is given by the formula below:

$$A_i^0 = \{\underline{x} : \min_i (f_M \cdot c_i^M \cdot \pi_0 + f_U \cdot c_i^U \cdot (1 - \pi_0))\} \quad (9)$$

In order for our model to define the decision areas, it makes use of n systems of $n - 1$ linear inequalities. By solving for the likelihood ratio f_M / f_U in each one of these systems:

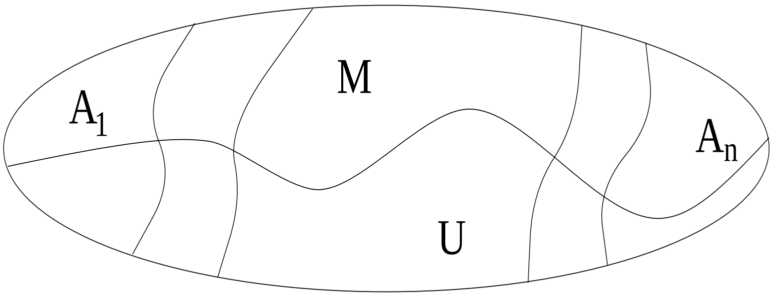

Figure 1: A partitioning of the decision space.

$$
\begin{aligned}
f_M/f_U &\leq (c_1^U - c_i^U)/(c_i^M - c_1^M) \cdot (1-\pi_0)/\pi_0 \\
&\vdots \\
f_M/f_U &\leq (c_{i-1}^U - c_i^U)/(c_i^M - c_{i-1}^M) \cdot (1-\pi_0)/\pi_0 \\
f_M/f_U &\geq (c_{i+1}^U - c_i^U)/(c_i^M - c_{i+1}^M) \cdot (1-\pi_0)/\pi_0 \\
&\vdots \\
f_M/f_U &\geq (c_n^U - c_i^U)/(c_i^M - c_n^M) \cdot (1-\pi_0)/\pi_0 \qquad (10)
\end{aligned}
$$

we get $n(n-1)$ values the likelihood ratio should be compared with. These values denote the thresholds that explicitly define the decision areas. By inspecting these values closely, we observe that half of them are unique. Notice for example, that the last inequality in A_1^o and the first in A_n^o give raise to $f_M/f_U \geq (c_n^U - c_1^U)/(c_1^M - c_n^M) \cdot (1-\pi_0)/\pi_0$ and $f_M/f_U \leq (c_1^U - c_n^U)/(c_n^M - c_1^M) \cdot (1-\pi_0)/\pi_0$ correspondingly, where the thresholds are exactly the same. In general, the n systems of $n-1$ equations generate $\binom{n}{2}$ unique thresholds. In order for all of the n decision areas to exist, the following sufficient and necessary condition should hold for $n-1$ of these thresholds:

$$
\begin{aligned}
&\frac{c_n^U - c_{n-1}^U}{c_{n-1}^M - c_n^M} \leq \frac{c_{n-1}^U - c_{n-2}^U}{c_{n-2}^M - c_{n-1}^M} \leq \dots \\
&\leq \frac{c_4^U - c_3^U}{c_3^M - c_4^M} \leq \frac{c_3^U - c_2^U}{c_2^M - c_3^M} \leq \frac{c_2^U - c_1^U}{c_1^M - c_2^M} \qquad (11)
\end{aligned}
$$

Notice that for simplicity reasons, the ratio of prior probabilities have been eliminated from all the thresholds. For example, if for the likelihood ratio of a comparison vector the following inequality holds:

$$
\frac{c_4^U - c_3^U}{c_3^M - c_4^M} \cdot \frac{1-\pi_0}{\pi_0} \leq \frac{f_M}{f_U} \leq \frac{c_3^U - c_2^U}{c_2^M - c_3^M} \cdot \frac{1-\pi_0}{\pi_0} \qquad (12)
$$

then the comparison vector belongs to A_3^o.

3.1 Optimality of the Decision Model

We can now prove that the decision model that we have proposed (i.e., the sets A_1^o, A_2^o, ..., A_n^o) is an optimal one. Based on the discussion above we know that

$$
A = A_1 \bigcup A_2 \bigcup \cdots \bigcup A_n,
$$

where A_1, A_2, $\cdots$, A_n are pair-wise disjoint. Every point will be assigned to either one of these decision areas. We also introduce the indicator function I_C of a set C, as the function which takes the value of 1 if the point $\underline{x}$ belongs to C and the value 0, otherwise. Note that we can formally write Eq. 8 as:

$$
\overline{c} = \sum_{\underline{x} \in A_1} z_1(\underline{x}) + \sum_{\underline{x} \in A_2} z_2(\underline{x}) + \cdots + \sum_{\underline{x} \in A_n} z_n(\underline{x}) \qquad (13)
$$

where $z_i(x)$, $i = 1, 2, \ldots, n$ denote the expressions inside the corresponding sums in Eq. 8.

Using the indicator functions, we can write:

$$
\begin{aligned}
\overline{c} &= \sum_{\underline{x} \in A_1} z_1(\underline{x}) + \sum_{\underline{x} \in A_2} z_2(\underline{x}) + \cdots + \sum_{\underline{x} \in A_n} z_n(\underline{x}) \\
&= \sum_{\underline{x} \in A} [z_1(\underline{x}) \cdot I_{A_1}(\underline{x}) + z_2(\underline{x}) \cdot I_{A_2}(\underline{x}) + \\
&\qquad \cdots + z_n(\underline{x}) \cdot I_{A_n}(\underline{x})] \\
&\geq \sum_{\underline{x} \in A} \min\{z_1(\underline{x}), z_2(\underline{x}), \ldots, z_n(\underline{x})\} \\
&\stackrel{\text{def}}{=} \sum_{\underline{x} \in A_1^o} z_1(\underline{x}) + \sum_{\underline{x} \in A_2^o} z_2(\underline{x}) + \cdots + \sum_{\underline{x} \in A_n^o} z_n(\underline{x}).
\end{aligned}
$$

3.2 Error Estimation

The probability of errors can now be easily computed. There are two types of errors. The first one is called Type I error, and it occurs when a *non-link* action is taken although the two records are actually matched. The probability of this error can be estimated as follows:

$$
\begin{aligned}
P(d = A_n, r = M) &= P(d = A_n | r = M) \cdot P(r = M) \\
&= \pi_0 \cdot \sum_{\underline{x} \in A_n} f_M(\underline{x}). \qquad (14)
\end{aligned}
$$

The second type of error is called Type II error and it occurs when the *link* action is taken although the pair of records is actually non-matched. The probability of this error can be estimated as follows:

$$
\begin{aligned}
P(d = A_1, r = U) &= P(d = A_1 | r = U) \cdot P(r = U) \\
&= (1-\pi_0) \cdot \sum_{\underline{x} \in A_1} f_U(\underline{x}). \qquad (15)
\end{aligned}
$$

By computing these two types errors, we assume that all the other areas, in between these two, are not considered as definite decisions, and for this reason, we can use points assigned to them in either kind of error before further investigation.

4. APPLICATION

The previously presented model will be demonstrated in a file maintenance application, where the source data are lists of subscribers of two large magazine publishers. Table 2 shows tentative unit costs developed by the staff of the publishers on the basis of consideration of the character of the actions and the consequences of these actions. For example, based on the contents of this table, the cost c_2^M is \$0.41. A possible set of actions that should be taken for a record comparison pair is presented below:

- Treat the comparison pair as if it designated to the same individual of some population. This is equivalent to the “link” decision.
- Temporarily treat the comparison pair as a link but obtain additional information before classifying the pair as a link or a non-link.
- Take no action immediately but obtain additional information before classifying the pair as a link or non-link.
- Temporarily treat the pair as if it was associated with different individuals of the population, but obtain additional information before classifying the pair as link or non-link.
- Treat the pair as if it was associated with different individuals in the population (non-link).

Table 2: Tentative Unit Costs

	True Status	
Action	Match	Non-match
1	\$0.00	\$6.01
2	0.41	1.13
3	0.77	0.77
4	0.82	0.41
5	2.59	0.00

In order to delineate the decision areas, we need to start with the test given in Eq. 11. By using this test we can find out whether all the areas are well defined, and if so, which are these areas for each action. In this example, the number of actions, or else decision areas, is 5. So, intuitively, four thresholds (the four rightmost ones in Eq. 11) can be checked. By substituting the values of the costs from Table 2 in Eq. 11 we get:

$$0.232 \leq 7.2 \leq 1 \leq 11.902 \tag{16}$$

It is obvious that in Eq. 16 not all of the thresholds are in the right order. This means that not all areas (5 decision areas) are defined by these costs so in order to define them, we then need to consider the initial detailed model and the corresponding systems of equations. The system of inequalities for this application is depicted in Table 3.

Notice that the unique thresholds for f_M/f_U in Table 3 are the r_{ij}'s, since the thresholds in the lower diagonal system are the same as their diagonal images. Also notice that $r_{ii} = 0$. Observe that the following two systems of inequalities should hold in order for all of the five areas to be well defined:

$$r_{12} \geq r_{13} \quad r_{12} \geq r_{14} \quad r_{12} \geq r_{15} \tag{17}$$

$$r_{23} \geq r_{24} \quad r_{23} \geq r_{25} \tag{18}$$

$$r_{34} \geq r_{35} \tag{19}$$

and

$$r_{35} \geq r_{45} \quad r_{25} \geq r_{45} \quad r_{15} \geq r_{45} \tag{20}$$

$$r_{24} \geq r_{34} \quad r_{14} \geq r_{34} \tag{21}$$

$$r_{13} \geq r_{23} \tag{22}$$

Notice, for example, that the system of inequalities in Eq. 17 holds because the threshold in the cell(2,1) in Table 3 (diagonal image r_{12}), needs to be the maximum threshold in the first row, otherwise there will be a gap between the first and the second decision areas.

By combining the inequalities above, we verify that $r_{12} \geq r_{23} \geq r_{34} \geq r_{45}$. In our case by substituting the values of the unit costs to the original system of inequalities, we get:

$$\begin{array}{lllll}
r_{11} & r_{12} = 11.09 & r_{13} = 6.805 & r_{14} = 6.82 & r_{15} = 2.32 \\
 & r_{22} & r_{23} = 1 & r_{24} = 1.75 & r_{25} = 0.51 \\
 & & r_{33} & r_{34} = 7.2 & r_{35} = 0.42 \\
 & & & r_{44} & r_{45} = 0.232 \\
 & & & & r_{55}
\end{array}$$

By processing the information in the above system, we generate the decision areas:

$$\text{Area} = \begin{cases} 1 & \text{if } f_M/f_U \geq 11.09 \\ 2 & \text{if } 11.09 \geq f_M/f_U \geq 1.75 \\ 4 & \text{if } 1.75 \geq f_M/f_U \geq 0.232 \\ 5 & \text{if } 0.232 \geq f_M/f_U \end{cases}$$

Area 3 is not feasible, since there is no region in the real axis in which $f_M/f_U \leq 1.75$ and at the same time $f_M/f_U \geq 7.2$. Notice also that the thresholds given above should be scaled by the ratio of prior probabilities $(1 - \pi_0)/\pi_0$. In the next section, we elaborate on this issue.

5. EXPERIMENTS AND RESULTS

In order to validate and evaluate the proposed decision model, we have built an experimental evaluation system [12]. The evaluation system is built on top of a public domain system, the database generator [3], that automatically generates source data, with user-selected a-priori characteristics. The database generator allows us to perform controlled studies so as to establish the accuracy (or else the overall error), the percentage of comparison pairs which are assigned to the various decision areas and the overall cost of the record linkage process. The database generator provides a large number of parameters for selection such as the size of the generated database, the percentage of duplicate records in the database, and the percentage of the error in the duplicated records. Each one of the generated records, consists of the fields shown in Table 4. Some of the fields, as well, can be empty, affecting in this way the presence value. As it is reported in [3] the names were chosen randomly from a list of 63000 real names. The cities, the states and the zip codes (all from the USA) come from publicly available lists.

Table 3: A 5-by-5 system of inequalities for the file maintenance application.

$$\begin{array}{ccccc}
r_{11} & \geq \frac{c_2^U - c_1^U}{c_1^M - c_2^M} = r_{12} & \geq \frac{c_3^U - c_1^U}{c_1^M - c_3^M} = r_{13} & \geq \frac{c_4^U - c_1^U}{c_1^M - c_4^M} = r_{14} & \geq \frac{c_5^U - c_1^U}{c_1^M - c_5^M} = r_{15} \\
\leq \frac{c_1^U - c_2^U}{c_2^M - c_1^M} & r_{22} & \geq \frac{c_3^U - c_2^U}{c_2^M - c_3^M} = r_{23} & \geq \frac{c_4^U - c_2^U}{c_2^M - c_4^M} = r_{24} & \geq \frac{c_5^U - c_2^U}{c_2^M - c_5^M} = r_{25} \\
\leq \frac{c_1^U - c_3^U}{c_3^M - c_1^M} & \leq \frac{c_2^U - c_3^U}{c_3^M - c_2^M} & r_{33} & \geq \frac{c_4^U - c_3^U}{c_3^M - c_4^M} = r_{34} & \geq \frac{c_5^U - c_3^U}{c_3^M - c_5^M} = r_{35} \\
\leq \frac{c_1^U - c_4^U}{c_4^M - c_1^M} & \leq \frac{c_2^U - c_4^U}{c_4^M - c_2^M} & \leq \frac{c_3^U - c_4^U}{c_4^M - c_3^M} & r_{44} & \geq \frac{c_5^U - c_4^U}{c_4^M - c_5^M} = r_{45} \\
\leq \frac{c_1^U - c_5^U}{c_5^M - c_1^M} & \leq \frac{c_2^U - c_5^U}{c_5^M - c_2^M} & \leq \frac{c_3^U - c_5^U}{c_5^M - c_3^M} & \leq \frac{c_4^U - c_5^U}{c_5^M - c_4^M} & r_{55}
\end{array}$$

For each study, the evaluation system makes an external call to the database generator in order to generate two databases. The first database is used for training the decision model and the second database for testing the model. The training process includes the estimation of the required parameters by the decision model. Both databases are generated by using almost the same parameter settings. Only the number of records and the number of record clusters in each database can be different. A record cluster is a group of records in the same database that refers to the same person. All the records in the same cluster are considered as duplicates. The training and the test databases are used correspondingly for generating the training comparison space and the test comparison space. In this study, the comparison vector has binary components and for this reason the result of a comparison can either be 0 or 1.

Some of the options that are provided to the users of the experimental system, for the generation of the training and test comparison spaces, include: (a) the pre-conditioning of the database records, (b) the selection of the sorting keys to be used for sorting the original database records, (c) the functions to be used for the comparison of each record attribute, (d) the searching strategy along with its parameters if applicable, and (e) the thresholds for the decision model. For the pre-conditioning of the database records, we may select to convert all the characters to uppercase or lowercase, and compute the Soundex code of the last name. Any subset or part of the record fields can be used as a sorting key. Among the functions to be selected for comparing pairs of field values, the most frequently used are the Hamming distance for numerical attributes, and the edit distance [7], the n-grams [4], the Jaro distance [5], and the Smith-Waterman algorithm [8] for character string attributes. For the searching strategy, the experimental system currently supports the blocking and the sorted-neighborhood approach. In the sorted-neighborhood approach the window size to be used should also be provided as an input parameter to the system. The last part of the parameters that are required by the system include the threshold values, which delimit the various decision areas in the proposed model.

In the set of experiments that we present, we make use of a comparison space of $10,000$ comparison records with known true matching status, as the training set, and a set of $1,000,000$ records in the testing set. Notice that the size of the comparison space depends heavily on the searching technique used and is usually close to an order of magnitude larger than the number of actual database records compared. The estimated probabilities of presence and agreement are given in Table 4. These probabilities can be easily computed by using the information in the training comparison space, since the actual matching status is considered known. This is possible, because each database record has been assigned a cluster identifier by the database generator, which is used for the identification of the cluster that each record belongs to.

Table 4: Estimated probabilities of presence and agreement in the training comparison space.

	True Status			
	Match		Non-match	
Attribute	$\hat{p}_j$	$\hat{p}_j^*$	$\hat{q}_j$	$\hat{p}_j^*$
SSN	0.87	0.85	0.81	0.15
First Name	0.91	0.87	0.83	0.08
Middle Initial	0.76	0.64	0.93	0.05
Last Name	0.86	0.75	0.83	0.21
Street Number	0.90	0.57	0.81	0.10
Street Address	0.67	0.58	0.88	0.07
Apartment Number	0.45	0.47	0.89	0.05
City	0.56	0.59	0.91	0.12
State	0.78	0.81	0.86	0.16
Zip Code	0.89	0.91	0.92	0.06

The system also uses the costs of the various actions, in the decision process. Here, we make use of the costs presented in Table 2. In the experiments, we have generated pairs of training and testing record comparison sets with a variable size of cluster size. We have run many experiments in order to estimate the total cost of the linkage process for each testing comparison set by using a variable size of comparison fields from the Table 4. The results are shown in Table 5 and indicate that (a) the cost of the record linkage process decreases as the dimensionality of the comparison space increases and (b) for fixed dimensionality, there is no clear evidence whether the prior matching probability affects positively or negatively the total cost. The first observation is consistent with the intuition that more "reliable" comparison components help the model to minimize the total cost, while the second observation, is necessary so as our model

Table 5: Total cost of record linkage for 1,000,000 comparison records. The prior probability is estimated on a set of 10,000 comparison records.

	Number of Vector Components				
$\hat{\pi}_0$	1	2	4	8	10
0.500	$222,700.000	113,359.125	18,790.750	6.300	1.900
0.250	200,206.250	109,679.312	19,582.665	9.038	2.700
0.200	173,105.000	101,050.850	18,540.950	9.544	2.880
0.125	125,140.625	80,936.280	15,839.407	10.230	3.090
0.100	109,152.500	74,231.425	14,563.803	10.420	3.150

to be independent of the data and so – to the degree this is possible – unbiased. Other experiments performed, indicate that our model provides always the most cost efficient linkage.

6. CONCLUSIONS

This paper presents a new cost optimal decision model for the record matching process. The proposed model uses the ratio of the prior odds along with appropriate values of thresholds to partition the decision space to a number of decision areas. The major difference between our model and the other existing models is that it minimizes the cost of making a decision rather than the probability of an erroneous decision. Our model is also much more efficient than other error-based models, as it does not resort to the sorting of the posterior odds in order to select the threshold values. The applicability of this model is independent of the characteristics of the comparison fields, of the database fields, of the sorting techniques used and of the matching functions.

In our future endeavors, we are also considering the design of a model for cost and time optimal record matching. By using such a model, it will be feasible not only to make a decision based on the entire comparison vector, but also to acquire as many comparison components as required, in order to make a certain decision. This will save computation time and at the same time it will facilitate the on-line decision making in the record matching context.

7. REFERENCES

[1] Wendy Alvey and Bettye Jamerson, *Record Linkage Techniques – 1997*, Proceedings of an International Workshop and Exposition, March 1997, Federal Committee on Statistical Methodology, Office of Management and Budget.

[2] I. P. Fellegi and A. B. Sunter, *A Theory For Record Linkage*, Journal of the American Statistical Association **64** (1969), no. 328, 1183–1210.

[3] Mauricio Antonio Harnández-Sherrington, *A Generalization of Band Joins and the Merge/Purge Problem*, Ph.D. thesis, Department of Computer Sciences, Columbia University, 1996.

[4] Jeremy A. Hylton, *Identifying and Merging Related Bibliographic Records*, Master's thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 1996.

[5] Matthew A. Jaro, *Advances in Record-Linkage Methodology as Applied to Matching the* 1985 *Census of Tampa, Florida*, Journal of the American Statistical Association **84** (1989), no. 406, 414–420.

[6] Beth Kliss and Wendy Alvey, *Record Linkage Techniques – 1985*, Proceedings of the Workshop on Exact Matching Methodologies, May 1985, Department of the Treasury, Internal Revenue Service, Statistics Income Division.

[7] U. Manber, *Introduction to Algorithms*, Addison-Wesley Publishing Company, 1989.

[8] Alvaro Edmundo Monge, *Adaptive Detection of Approximately Duplicate Records and the Database Integration Approach to Information Discovery*, Ph.D. thesis, Department of Computer Science and Engineering, University of California, San Diego, 1997.

[9] H.B. Newcombe and J.M. Kennedy, *Record Linkage: Making Maximum Use of the Discriminating Power of Identifying Information*, Communications of the ACM **5** (1962), 563–566.

[10] H.B. Newcombe, J.M. Kennedy, S.J. Axford, and A.P. James, *Automatic Linkage of Vital Records*, Science **130** (1959), no. 3381, 954–959.

[11] Benjamin J. Tepping, *A Model for Optimum Linkage of Records*, Journal of the American Statistical Association **63** (1968), 1321–1332.

[12] Vassilios S. Verykios, Mohamed G. Elfeky, Ahmed K. Elmagarmid, Munir Cochinwala, and Sid Dalal, *On the Accuracy and Completeness of the Record Matching Process*, 2000 Information Quality Conference (2000), 54–69.

[13] Vassilios S. Verykios, George V. Moustakides, and Mohamed G. Elfeky, *A Bayesian Decision Model for Cost Optimal Record Matching*, The VLDB Journal **12** (2003), no. 1, 28–40.