Selecting potentially relevant records using re-identification methods

Josep Domingo-Ferrer[1] and Vicenç Torra[2]

[1] *Dept. Comput. Eng. and Maths - ETSE, Universitat Rovira i Virgili, Av Països Catalans 26, 43007 Tarragona (Catalonia, Spain)*

[2] *Institut d'Investigació en Intel·ligència Artificial - CSIC, Campus UAB s/n, 08193 Bellaterra (Catalonia, Spain)*

jdomingo@etse.urv.es

vtorra@iiia.csic.es

Abstract This work proposes re-identification algorithms for selecting records or variables that are interesting from the point of view of giving new information. Instead of focusing on re-identified elements (records or variables), we focus on non re-identified ones (non-linked elements) as they are the ones that potentially supply new and relevant information. Moreover, these relevant characteristics can correspond to chances for improving the knowledge of a system.

Our approach has been empirically validated by applying it to an example using publicly available data from the UCI repository. We have used the data of the *ionosphere* database to build a re-identification problem for 35 non-common variables.

Keywords Chance discovery, Knowledge discovery in databases, Data mining, Multi-database mining, Re-identification algorithms, Record selection, Record linkage, Variable correspondence identification.

§1 Introduction

Knowledge discovery in databases (KDD) means deriving models for large datasets (large files with a great number of records and variables). According to [1)], the discovery goal can be divided into a *descriptive* part and a

predictive part. The descriptive part is about finding models for the data in a way that they become understandable to the user. The predictive part is about finding models for the data so that the future behavior of a (set of) variable(s) for a particular individual is established. Data mining techniques (techniques for building particular models from data) are designed to achieve knowledge discovery.

The so-called re-identification methods are a specific class of data mining methods. These methods are either designed to relate data from different files (or databases) that correspond to the same individuals or to relate variables from different files that correspond to the same context. Making relationships explicit makes sense at least in the following situations:

1. Data concerning individuals (*e.g.* persons, companies, ...) tend to be nowadays scattered across several organizations (and often, across several departments within an organization) To access data in a consistent way, records corresponding to the same individual should be related. Application of complex tools to these databases requires data to be consistent (entity integrity has to be satisfied)[*1]. Record and variable identification methods are applied for this purpose.
2. Relationships between objects are a kind of knowledge that can be exploited in certain environments. Re-identification applied to files with non-common individuals allow similarities to be established between objects that have similar properties or behaviors.
3. Records for which relationships cannot be established give potentiality to knowledge discovery applications. Failure to find relationships between a record and a set of individuals can either be due to relevant modification of the knowledge related to a particular individual or to a completely new piece of knowledge. In the case of variables, failure can be due to significant information updating (information in year 2003 can substantially differ from information in year 1993) or to some of the variables being really different/new. Thus, non-linked records or variables are interesting because correspond to elements of potentially relevant characteristics and, therefore, chances exist [2)] at this point for improving the knowledge of a system.

[*1] It should be noted that often this is not the case. An example is advertisement mailings: how many times have we received duplicated letters from the same company due to inconsistent handling of names and addresses?

In this work, we focus on the last issue. Our approach follows [3] and their study of the role of counterexamples in discovery learning environments. As they point out:

> In a learning environment, a learner tries her/his solution and often makes mistakes. The occurrence of such errors becomes a good opportunity for learning, that is, a chance. Novel phenomena are often given as counterexamples, that indicate the difference between her/his prediction and the result. They have potential to cause a learner's conceptual change.

From our point of view, non-linked elements in our setting are analogous to the counterexamples in the work by [3]. This is so because we understand that in our case the usual prediction is to link records with the model (the available data). Therefore, non-linked records deviate from the prediction and, accordingly, they have potential to cause some correction on the initial model (the original data file).

Detection of new and potentially interesting knowledge is of interest for re-identification purposes. Nevertheless, it must be said that this is not a complete process. Although several methods have been defined for re-identification, current technology [4] does not allow a complete automatic translation and a certain amount of clerical work is required. Therefore, a high degree of human-computer interaction is needed so that the system proposed helps the user to detect opportunities.

We would like to underline that, although the description is given considering the re-identification of sets of records in a data file, the same approach can be applied to the re-identification of variables or to any other relational structure (*e.g.* datafiles or predicates). In general, we can consider any data model. In our case, we assume that the initial model (that can be revised or modified) simply corresponds to a set of records; however, other (more complex) models are also be conceivable.

We consider in what follows several types of non-linked records. That is, we consider several types of failure for the comparison between two records corresponding to the same individual.

Modification of the initial model: Failure of the re-identification procedure is caused by modification of the value for a relevant variable. Thus, the value that the individual gave to the same variable in the past (and which was recorded in the initial model) has changed in a meaningful

way. Following the analogy with the work by Horiguchi and Hirashima [3], this would correspond to a discontinuity of thinking.

An example of this situation is when the model contains information about companies and an old company is replaced by a completely new company (which keeps most of the customers and providers of the old company). A simpler (and easier to detect) case would be a change of the corporate address of the company. In this case, detection of partial matching would be possible.

New information: Failure corresponds to the appearance of a new individual. No previous information about that individual was known. The record does not match any other records because no information on the former was included in the initial model.

No new information: The non-linked record corresponds to an individual (either new or a modification of a previous one) quite similar to existing records. Therefore, no new information is available. This is the case if a cluster of companies with similar characteristics (similar production, providers, customers) is considered and a new similar company appears in the picture.

In this work, we propose heuristic methods for detecting the three aforementioned different types of non-linked records. Our approach is based on re-identification algorithms. The structure of this paper is as follows. Section 2 contains a short description of re-identification algorithms. Section 3 presents our approach. Section 4 lists some conclusions and some ideas for future work.

§2 Re-identification methods

The approach presented here is based on re-identification algorithms. As pointed out before, these are methods to link records in different data files that correspond to the same individuals. Re-identification is not an easy task. The usual assumptions and problems for developing such methods are described below. In general, two different situations can be distinguished depending on whether the files share a set of common variables:

1. When files share a set of common variables, it is often the case that records that correspond to the same individual do not completely match. As pointed out by Winkler[5] "the normal situation in record linkage is that identifiers in pairs of records that are truly matches disagree by small or large amounts and that different combinations of the non-

unique, error-filled identifiers need to be used in correctly matching different pairs of records". This fact is mainly due to noisy data (either due to intentional or accidental noise [6]).

2. When files do not share a set of common variables (but correspond to the same set of individuals), finding relationships between individuals is more difficult. In this case, it is not possible to compare one-by-one records in one file with records in the other file due to the lack of common variables. To make some re-identification possible (see e.g. [7]), a common assumption is that both files contain, implicitly, similar structural information. Differences between methods stem from differences in the way the structural information is conceived (*e.g.* partitions or dendrograms). Re-identification is then based on the assumption that similarities within one file are maintained in the other file; this assumption is not unreasonable, because the set of individuals described by both files is the same.

In the rest of this section, we review some re-identification algorithms. We start with those designed for files that share a set of variables and then we discuss those that do not require such an assumption.

2.1 Re-identification with common variables

In the case of common variables, two main approaches are used in the literature: probabilistic record linkage and distance-based record linkage. We consider in both cases the re-identification of the records of a file B with respect to a file A. Files A and B share a set of common variables. It would not be realistic to assume that the values in both files are the same for the same individuals. That is, if we take the same variable and the records corresponding to the same individual in both files, it is often the case that the values we get are not exactly the same.

[1] Probabilistic record linkage

In this section, we review some of the main characteristics of these methods based on [8]. Detailed descriptions of these methods are given in [5], [9] and [10].

We start considering files A and B with a single variable v in each file. Then, let r_A and r_B be records from files A and B. Also, let us consider that the values for records r_A and r_B for the variable v are a and b, respectively. That

is, $r_A = a$ and $r_B = b$.

Probabilistic record linkage is based on the computation of an index for each pair (r_A, r_B). According to this index and certain thresholds, the pair is classified either as a linked pair (LP), a clerical pair (CP) or a non-linked pair (NP). The index for the pair (r_A, r_B) is denoted by $R(a, b)$ and is computed as follows:

$$R(a,b) = log\Big(\frac{P(a = b|(a,b) \in \mathbf{M})}{P(a = b|(a,b) \in \mathbf{U})}\Big) \quad (1)$$

where $\mathbf{M}$ corresponds to the set of *matched pairs* and $\mathbf{U}$ corresponds to the set of *unmatched pairs*. That is, $\mathbf{M}$ are the pairs that can be proven to be real matches (with both records corresponding to the same individual) and $\mathbf{U}$ are the pairs that can be proven to be unrelated (with records in the pair corresponding to different individuals).

When both files contain a set of variables rather than a single variable, an expression equivalent to Expression (1) is used. This case requires a and b to be vectors of values for a set of variables V instead of single values for the given variable v. In order to compute $R(a, b)$, a common assumption is that different variables are statististically independent and, thus, products of conditional probabilities can be used. Winkler discusses in [11] alternative approaches not assuming conditional independence between variables in V.

The application of probabilistic record linkage requires setting the thresholds for determining whether a pair is a match (say, the values *linkThreshold* and *nonLinkThreshold*) and the conditional probabilities in Expression (1).

- The thresholds are usually determined from the probabilities:

  $$P(LP|\mathbf{U})$$

  $$P(NP|\mathbf{M})$$

  This is:

  (i) the probability of linking a pair that is an unmatched pair (a *false positive* or *false linkage*)

  (ii) the probability of not linking a pair that is a match pair (a *false negative* or *false unlinkage*).

- The conditional probabilities in Expression (1) are usually estimated using the EM algorithm [12] when no information is given about the sets $\mathbf{M}$ and $\mathbf{U}$. If some examples of pairs of records belonging to both sets are

available, they can be used to infer those conditional probabilities. See [13] for details on probabilistic record linkage.

[2] Distance-based record linkage

This approach, described in [14] in a very restricted formulation and described in more detail in [15], consists of computing the distances between records in the two data files under consideration.

In general, this method is as follows: i) for each record in file A, the distance to every record in file B is computed; ii) then the *nearest* record in file B is considered. A record in file B is correctly linked when the nearest record in file A turns out to be its corresponding original record (nonzero distance between corresponding records means that one of the records has received some distortion). In all other cases, records are not correctly linked.

The distance-based approach requires that distances be standardized to avoid scaling problems. Also, some assumption is required on the weights of variables when computing distances (the same weight for all variables was used in [14] and [15]).

[3] Discussion

Both record linkage methods try to find the records in files A and B that correspond to the same individuals. As shown above, both approaches are radically different. Following [8], we underline the following aspects:

- Distance-based record linkage methods are simple to implement and operate. The main difficulty for using them is choosing appropriate distances for the variables under consideration. In particular, distances for categorical variables (in ordinal and nominal scales) are required. On the other hand, distance-based record linkage allows the inclusion of additional information (either subjective information or domain knowledge) about individuals or variables in the re-identification process.
- Probabilistic record linkage methods are less simple. On the positive side, they do not need scaling or weighting of variables (weights are determined by the EM algorithm) and require the user to provide only two probabilities as input: these are $P(LP|\mathbf{U})$ and $P(NP|\mathbf{M})$.

Furthermore, our comparative analyses show that both methods, although built to achieve the same goals, do not lead to the same results. See [15] and [8] for comparison of both methods when variables are, respectively, numer-

ical and categorical.

2.2 Re-identification without common variables

A different situation for re-identification is the case where files do not share any variables. In such a case, the above mentioned methods are not applicable unless some previous transformations are made to the data.

In general, re-identification is achievable if the underlying structure of both files is similar. Different approaches arise depending on the way of extracting structural information. Once this information is extracted, re-identification is performed based on it.

A first approximation to re-identification without common variables was reported in [7], where the underlying structure in both files was expressed by means of partitions. Several clustering algorithms were applied to both files, and a partition of the records was obtained with each algorithm. Then, re-identification was applied at the level of partitions: records clustered together in one file should correspond to records clustered together in the other file. A similar approach is described in [16].

A different approach was considered in [17]. In this work, several prototypes were computed for each record. On the basis that prototypes should be similar in both files, re-identification was applied. Prototypes were obtained using OWA operators [18] - a parametric aggregation operator.

§3 Selection of potentially interesting records

As stated in the introduction, the failure of re-identification algorithms can signal the presence of potentially interesting records. According to this, additional work is needed to find the causes of such a failure and select those records that represent some new information to be taken into account by the system. In some sense, non-linked records represent chances of finding relevant knowledge and, following [19], what we try here is to become aware of these chances and explain their significance.

In short, we rely on the following hypothesis:

Hypothesis 1
Re-identification algorithms can discover, on one hand, new relationships between objects (by finding pairs of linked records) and, on the other hand, chances of finding potentially interesting records (by yielding non-linked records).

In fact, the first part of this assumption is obviously true, so we focus on whether it is possible to find potentially interesting records by means of record linkage. To that end, we introduce a methodology to further analyze the set of non-linked records obtained after applying some re-identification algorithms. In the remainder of this section, we consider a single non-linked record belonging and we try to classify it in one of the three types pointed out in the introduction. That is, either the record: (i) corresponds to knowledge already included in the original file (in the initial model) but some relevant information is updated with respect to the original file; (ii) corresponds to (completely) new information; or (iii) does not provide any new information.

In order to decide whether a record r belongs to the first group, we need to match the record with the full original file without considering the updated variables W (*i.e.*, matching restricted to a subset of the variables). In this case, two alternatives can be distinguished:

(i) When the output of the matching procedure is a set S consisting of one or a few records, it may happen that the record r that should be re-identified belongs to S.

(ii) When the output of the matching procedures is set S of records with large cardinality, we have that the reduction of information caused by the reduction of the number of variables completely protects the record r from re-identification. That is, if the record was already present in the original file, the modification of relevant characteristics prevents it from being found.

This process can be formulated in the following rule:

Rule for detecting a modification of the initial model:

`If` $matches(r, A, V) = \emptyset$ `and`

$cardinality(matches(r, A, V - W)) \geq 1$ `and`

$cardinality(matches(r, A, V - W)) = small_{matches}$

`then`

r possibly matches elements in S

where $S = matches(r, A, V - W)$ and $small_{matches}$ is a *fuzzy* term.

Here, *matches* corresponds to the application of the original re-identification algorithm to the record under study and the original file A. This function returns the set of re-identified records, *i.e.* $matches(r, A, V)$ corresponds to the records that match with r in the file A when considering variables in V. Therefore, the third parameter (either V or $V - W$ in the rule above) corresponds to

the variables to be considered in the matching procedure.

Note that given a record r, it is usual that $S_W = matches(r, A, V - W)$ is different for different sets of variables W. In this case, among all the possible sets W, the best ones are the ones with a small cardinality. This means to remove as few variables as possible. In fact, in general, the best selection for W is the set that leads to a minimum (nonzero) cardinality of S_W.

In order to direct the search and avoid considering all subsets W of V, we propose to bear in mind the approaches given below for selecting the subset W. These approaches are based on the knowledge available on the variables.

Reliability of the variables: We consider as reliable those variables whose values are unlikely to change. Unreliable variables have to be considered first in W.

Variability of the variables: There is variability when small errors (or other causes) can provoke small variations on the data. Variability can make re-identification fail. Variables with large variability have to be considered first.

Weights of the variables: Some re-identification methods (*e.g.* probabilistic methods and the EM algorithm) allow a weight to be computed for each variable so that its influence in the re-identification is explicitly quantified. Variables with large weights should be considered first.

Nearest record (according to any of the re-identification methods): The comparison of a record with the nearest one signals values that are different. The corresponding variables can be taken into account.

If the record does not correspond to an updating of some existing information, the results of the matching process with a reduced number of variables give information for selecting any of the other types of non-linked records:

New information: If no set of variables W lead to some re-identification, then it means that the new record is some new information not present in the A file. This is also the case when re-identifications are only found for very small set of variables W. This can result in some false re-identification (note that the smaller the number of variables, the more likely is to find a record that exactly matches r).

No new information: This is when the record belongs to a cluster in the original file and most of the information is already in the original file A even though the particular record r is not in A.

Rule for detecting a modification of the initial model:
If $matches(r, A, V) = \emptyset$ and
  $cardinality(matches(r, A, V - W)) \geq 1$ and
  $cardinality(matches(r, A, V - W)) = small_{matches}$
then
  r possibly matches elements in S
Rule for detecting some new information:
If $matches(r, A, V) = \emptyset$ and
  $cardinality(matches(r, A, V - W)) = \emptyset$
    for all W such that $cardinality(W) = small_{numberOfVariables}$
then
  r is some new information
Rule for detecting a case of no new information:
If $matches(r, A, V) = \emptyset$ and
  $belongs(r, nearestCluster(r, A, V))$
then
  r does not contain additional information

Fig. 1 Heuristic classification rules for non-linked records

To sum up, Figure 1 contains the heuristic rules to classify the three situations we have considered in this section: (i) modification of the initial model; (ii) new information and (iii) no new information. Note that these rules are heuristic and thus, they only give hints about relevance of records. Also, the limits of the types of non-linked records are fuzzy and this is also the case for these rules.

3.1 Case study

In this subsection, we outline the application of the approach described above to a problem of re-identification of variables. This problem has been recently considered in the literature by several authors under several assumptions.

We have applied our approach to the re-identification example for non-common variables described in [7]. This example, based on publicly available data from the UCI repository [20], consisted on the re-identification of the 34 numerical variables (plus the classification variable) of the *ionosphere* database using as the original data 175 of the 351 examples (randomly selected) and as the re-identifiable data the remaining examples.

Table 1 Records as contained in the initial model.

		aa,m	aa,d	aa,t	cc,m	cc,d	cc,t
1	O_1^B	$c_{1,1}$	$c_{2,4}$	$c_{3,1}$	$c_{4,3}$	$c_{5,5}$	$c_{6,3}$
2	O_2^B	$c_{1,4}$	$c_{2,2}$	$c_{3,4}$	$c_{4,4}$	$c_{5,5}$	$c_{6,5}$
3	$O_4^B O_6^B$	$c_{1,2}$	$c_{2,2}$	$c_{3,3}$	$c_{4,1}$	$c_{5,2}$	$c_{6,2}$
4	$O_5^B O_7^B$	$c_{1,1}$	$c_{2,2}$	$c_{3,1}$	$c_{4,2}$	$c_{5,2}$	$c_{6,1}$
5	O_9^B	$c_{1,1}$	$c_{2,2}$	$c_{3,1}$	$c_{4,2}$	$c_{5,1}$	$c_{6,1}$
6	$O_8^B O_{10}^B$	$c_{1,2}$	$c_{2,2}$	$c_{3,3}$	$c_{4,1}$	$c_{5,1}$	$c_{6,2}$
7	$O_3^B O_{11}^B$	$c_{1,1}$	$c_{2,2}$	$c_{3,1}$	$c_{4,2}$	$c_{5,5}$	$c_{6,1}$
8	O_{12}^B	$c_{1,2}$	$c_{2,2}$	$c_{3,3}$	$c_{4,1}$	$c_{5,7}$	$c_{6,2}$
9	O_{13}^B	$c_{1,1}$	$c_{2,2}$	$c_{3,1}$	$c_{4,2}$	$c_{5,3}$	$c_{6,1}$
10	O_{14}^B	$c_{1,2}$	$c_{2,1}$	$c_{3,3}$	$c_{4,1}$	$c_{5,3}$	$c_{6,2}$
11	O_{16}^B	$c_{1,2}$	$c_{2,1}$	$c_{3,3}$	$c_{4,1}$	$c_{5,1}$	$c_{6,2}$
12	O_{20}^B	$c_{1,3}$	$c_{2,1}$	$c_{3,3}$	$c_{4,1}$	$c_{5,3}$	$c_{6,2}$
13	$O_{18}^B O_{22}^B$	$c_{1,3}$	$c_{2,1}$	$c_{3,3}$	$c_{4,1}$	$c_{5,2}$	$c_{6,2}$
14	$O_{15}^B O_{19}^B$						
	O_{23}^B	$c_{1,1}$	$c_{2,1}$	$c_{3,1}$	$c_{4,2}$	$c_{5,3}$	$c_{6,1}$
15	O_{24}^B	$c_{1,3}$	$c_{2,1}$	$c_{3,3}$	$c_{4,1}$	$c_{5,4}$	$c_{6,2}$
16	O_{26}^B	$c_{1,3}$	$c_{2,1}$	$c_{3,3}$	$c_{4,1}$	$c_{5,6}$	$c_{6,2}$
17	O_{27}^B	$c_{1,1}$	$c_{2,1}$	$c_{3,1}$	$c_{4,2}$	$c_{5,6}$	$c_{6,1}$
18	O_{28}^B	$c_{1,3}$	$c_{2,1}$	$c_{3,3}$	$c_{4,1}$	$c_{5,1}$	$c_{6,2}$
19	$O_{17}^B O_{29}^B$	$c_{1,1}$	$c_{2,1}$	$c_{3,1}$	$c_{4,2}$	$c_{5,1}$	$c_{6,1}$
20	$O_{25}^B O_{31}^B$	$c_{1,1}$	$c_{2,1}$	$c_{3,1}$	$c_{4,2}$	$c_{5,4}$	$c_{6,1}$
21	$O_{21}^B O_{33}^B$	$c_{1,1}$	$c_{2,1}$	$c_{3,1}$	$c_{4,2}$	$c_{5,2}$	$c_{6,1}$
22	$O_{30}^B O_{32}^B$						
	O_{34}^B	$c_{1,3}$	$c_{2,3}$	$c_{3,3}$	$c_{4,1}$	$c_{5,2}$	$c_{6,2}$
23	O_{35}^B	$c_{1,1}$	$c_{2,3}$	$c_{3,2}$	$c_{4,3}$	$c_{5,2}$	$c_{6,4}$

From a database perspective, the data files used are not large. Nevertheless, a larger number of records would increase the redundancy of the data and would make re-identification easier.

In [7], the original re-identification problem without common variables was translated into a re-identification problem with 6 common variables by means of 6 different clustering methods (see Section 2.2). Here, we start with the data once the problem has been translated into a problem with common variables. Therefore, at this point, re-identification methods for common variables can be applied. In particular, we have used distance-based record linkage in combination with a mapping to relate the ranges of the 6 variables.

Here we reproduce in Tables 1 and 2 the data to be re-identified. Table 1 lists the file considered as the initial model. This file contains the values for the 6 variables. The first column of the table corresponds to an index of the record, the second corresponds to those objects with the same evaluation for all 6 variables, *i.e.* indistinguishable objects on the basis of those 6 variables. The last six columns correspond to the values for the 6 variables. In these columns, aa and cc correspond to the classification criteria and stand for arithmetic average and centroid clustering, respectively; m, d and t refer to similarity functions based, respectively, on Manhattan distance, differences and taxonomic distance. $c_{\alpha,\beta}$ denotes the β-th partition of the α-th variable.

Table 2 lists the elements to be re-identified, their values for the 6 variables and to which records in Table 1 they are linked (column $r.i.$ stands for re-identification index). As in Table 1, all records with the same values are put together (list of objects in the first column). Re-identification shows that, in this example, 7 groups of records are not re-identified. These groups are listed in Table 3. For each group, only one of the records is given in this table. Analysis of the set of non-reidentified groups of records shows that by removing a single variable we can obtain exact links for all groups. As the best set W is, according to our approach, the one with the smallest number of variables, we only consider sets consisting of a single variable. Table 3 lists the linked groups of Table 1 for each removed variable. For example, value 18 in position aa, d of record O^A_{32} indicates that, when removing the variable aa, d, the record O^A_{32} is re-identified with objects in position 18 in Table 1 (*i.e.*, object O^B_{28}).

For some of the groups (those containing records $O^A_{10}, O^A_{29}, O^A_{33}, O^A_{34}, O^A_{35}$), there is only one of the variables that can be removed to obtain a non-empty set of linked records. Therefore, in this case, it is appropriate to consider that

the records only correspond to modifications of the initial model for a certain variable and the variable is now known. Moreover, analysis of the original data confirms that this hypothesis is correct (see, *e.g.* that records 10 and 11 in Table 1 only differ on variable cc, d from O^A_{10} in Table 2).

For the record O^A_{32}, there are two variables that lead to non-empty sets of linked records. From Table 1, we see that these records are singletons. For the group containing the record O^A_{21}, there are two variables that lead to non empty sets of linked records: variables aa, d and cc, d. If we consider the cardinality of the resulting sets, it seems appropriate to define $W = \{aa, d\}$ (*i.e.* define the selected set of variables as $\{aa, d\}$) because the linked set is a singleton (this would not be the case if $W = \{cc, d\}$). However, a detailed analysis of the variables points out that the variable cc, d has a large variability. This is so because all difficulties with the re-identification were due to this latter variable (compare the values for this variable in both Tables 1 and 2). According to this, and following our statement in Section 3 about variability, we can define $W = \{cc, d\}$ for both O^A_{21} and O^A_{32}.

We can conclude that record O^A_{32} seems to be a case of a modification of the initial model and record O^A_{21} seems either a case of modification of the initial model or of no new information. In this case, the new record O^A_{21} is near to 6 other groups of records in the original file.

§4 Conclusions and future work

In this work we have studied the selection of potentially relevant records. We have considered a data file and compared it against a previously existing data model (another data file in our case). This comparison has been achieved using re-identification methods. We have shown that the selection of potentially relevant records can be based on the non-linked records. As non-linked records point out chances of finding relevant knowledge, what we provide here is an awareness of these chances and an explanation of their significance.

This work introduces some heuristic rules for analyzing these potentially relevant records. Also, we have studied the results presented in [7] about identification of variables in the light of our approach. This analysis shows that it is also possible to relate records that correspond to real matches. Our recent work in [21] shows that the re-identification approach is general enough to be applied to larger sets and, thus, indicates that the methodology introduced here can also be applied for record linkage. For all these reasons, we consider that Hypothesis

Table 2 Records to be re-identified.

	r.i.	aa,m	aa,d	aa,t	cc,m	cc,d	cc,t
O_1^A	1	$c_{1,1}$	$c_{2,4}$	$c_{3,1}$	$c_{4,3}$	$c_{5,5}$	$c_{6,3}$
O_2^A	2	$c_{1,4}$	$c_{2,2}$	$c_{4,4}$	$c_{4,4}$	$c_{5,5}$	$c_{6,5}$
O_4^A	3	$c_{1,2}$	$c_{2,2}$	$c_{3,3}$	$c_{4,1}$	$c_{5,2}$	$c_{6,2}$
O_5^A	4	$c_{1,1}$	$c_{2,2}$	$c_{3,1}$	$c_{4,2}$	$c_{5,2}$	$c_{6,1}$
$O_6^A O_8^A$	6	$c_{1,2}$	$c_{2,2}$	$c_{3,3}$	$c_{4,1}$	$c_{5,1}$	$c_{6,2}$
$O_7^A O_9^A$	5	$c_{1,1}$	$c_{2,2}$	$c_{3,1}$	$c_{4,2}$	$c_{5,1}$	$c_{6,1}$
O_{10}^A	10	$c_{1,2}$	$c_{2,1}$	$c_{3,3}$	$c_{4,1}$	$c_{5,5}$	$c_{6,2}$
$O_3^A O_{11}^A$	7	$c_{1,1}$	$c_{2,2}$	$c_{3,1}$	$c_{4,2}$	$c_{5,5}$	$c_{6,1}$
O_{14}^A	10	$c_{1,2}$	$c_{2,1}$	$c_{3,3}$	$c_{4,1}$	$c_{5,3}$	$c_{6,2}$
O_{15}^A	14	$c_{1,1}$	$c_{2,1}$	$c_{3,1}$	$c_{4,2}$	$c_{5,3}$	$c_{6,1}$
$O_{12}^A O_{16}^A$	11	$c_{1,2}$	$c_{2,1}$	$c_{3,3}$	$c_{4,1}$	$c_{5,1}$	$c_{6,2}$
O_{18}^A	16	$c_{1,3}$	$c_{2,1}$	$c_{3,3}$	$c_{4,1}$	$c_{5,6}$	$c_{6,2}$
O_{19}^A	17	$c_{1,1}$	$c_{2,1}$	$c_{3,1}$	$c_{4,2}$	$c_{5,6}$	$c_{6,1}$
O_{20}^A	15	$c_{1,3}$	$c_{2,1}$	$c_{3,3}$	$c_{4,1}$	$c_{5,4}$	$c_{6,2}$
O_{21}^A	21	$c_{1,1}$	$c_{2,1}$	$c_{3,1}$	$c_{4,2}$	$c_{5,5}$	$c_{6,1}$
$O_{22}^A O_{24}^A$	18	$c_{1,3}$	$c_{2,1}$	$c_{3,3}$	$c_{4,1}$	$c_{5,1}$	$c_{6,2}$
O_{25}^A	20	$c_{1,1}$	$c_{2,1}$	$c_{3,1}$	$c_{4,2}$	$c_{5,4}$	$c_{6,1}$
$O_{13}^A O_{17}^A$							
$O_{23}^A O_{27}^A$	19	$c_{1,1}$	$c_{2,1}$	$c_{3,1}$	$c_{4,2}$	$c_{5,1}$	$c_{6,1}$
O_{28}^A	22	$c_{1,3}$	$c_{2,3}$	$c_{3,3}$	$c_{4,1}$	$c_{5,2}$	$c_{6,2}$
$O_{29}^A O_{31}^A$	19	$c_{1,1}$	$c_{2,3}$	$c_{3,1}$	$c_{4,2}$	$c_{5,1}$	$c_{6,1}$
$O_{26}^A O_{30}^A$							
O_{32}^A	22	$c_{1,3}$	$c_{2,3}$	$c_{3,3}$	$c_{4,1}$	$c_{5,1}$	$c_{6,2}$
O_{33}^A	21	$c_{1,1}$	$c_{2,3}$	$c_{3,1}$	$c_{4,2}$	$c_{5,2}$	$c_{6,1}$
O_{34}^A	22	$c_{1,3}$	$c_{2,3}$	$c_{3,3}$	$c_{4,1}$	$c_{5,7}$	$c_{6,2}$
O_{35}^A	23	$c_{1,1}$	$c_{2,3}$	$c_{3,2}$	$c_{4,3}$	$c_{5,1}$	$c_{6,4}$

Table 3 Re-identifications with $cardinality(W) = 1$

records	aa,m	aa,d	aa,t	cc,m	cc,d	cc,t
O_{10}^A	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	10, 11	$\emptyset$
O_{29}^A	$\emptyset$	5, 19	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$
O_{33}^A	$\emptyset$	4, 21	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$
O_{34}^A	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	22	$\emptyset$
O_{35}^A	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	25	$\emptyset$
O_{32}^A	$\emptyset$	18	$\emptyset$	$\emptyset$	22	$\emptyset$
O_{21}^A	$\emptyset$	7	$\emptyset$	$\emptyset$	14, 17, 19, 20, 21	$\emptyset$

1 presented in Section 3 has been validated.

Future extension of our work will be directed to obtaining new heuristic rules for selecting other kinds of relevant records. Also, we plan to apply the approach to larger datasets, for example sets from national statistical offices as in [15]. This would extend our work described in [21]. This application is would provide an experimental validation of the approach.

§5 Acknowledgments

Josep Domingo-Ferrer and Vicenç Torra are partially supported by the EU project CASC (IST-2000-25069) and by MCyT and the FEDER fund through project STREAMOBILE (TIC2001-0633-C03-01/02).

References

[1] Prendinger, H., Ishizuka, M., (2001), Methodological considerations on chance discovery, Lecture Notes in Computer Science (LNAI subseries), 2253, 425-434.

[2] Ohsawa, Y., Fukuda, H., (2000), Potential motivations as fountains of chances, Proc. of the IEEE Int. Conf. on Industrial Electronics, Control and Instrumentation (IECON 2000), 1626-1631.

[3] Horiguchi, T., Hirashima, T., (2001), The role of counterexamples in discovery learning environments: awareness of the chance for learning, Lecture Notes in Computer Science (LNAI subseries), 2253, 468-474.

[4] Rahm, E., Bernstein, P. A., (2001), A survey of approaches to automatic schema matching, VLDB Journal, 10, 334-350.

[5] Winkler, W. E., (1995), Matching and record linkage, in Business Survey Methods, B. G. Cox (Ed.), Wiley, 355-384.

[6] Domingo-Ferrer, J., Torra, V., (2001), Disclosure control methods and information loss for microdata, 91-110, in Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies, P. Doyle, J. I. Lane, J. J. M. Theeuwes, L. M. Zayatz (Eds.), Elsevier.

[7] Torra, V., (2000), Towards the re-identification of individuals in data files with non-common variables, Proc. of the European Conference on Artificial Intelligence, ECAI, 326-330, Berlin, Germany.

[8] Domingo-Ferrer, J., Torra, V., (2002), Validating distance-based record linkage with probabilistic record linkage, Lecture Notes in Computer Science (LNAI subseries), 2504, 207-215.

[9] Fellegi, I. P., Sunter, A. B., (1969), A theory for record linkage, Journal of the American Statistical Association, 64: 328, 1183-1210.

[10] Jaro, M. A., (1989), Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida, Journal of the American Statistical Association, 84, 414-420.

[11] Winkler, W. E., (1995), Advanced methods for record linkage, American Statistical Association, Proceedings of the Section on Survey Research Methods, 467-472.

[12] Dempster, A. P., Laird, N. M., Rubin, D. B., (1977), Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society, 39, 1-38.

[13] Gill, L., (2001), Methods for Automatic Record Matching and Linking and Their Use in National Statistics, London: Office for National Statistics.

[14] Pagliuca, D., Seri, G., (1999), Some Results of Individual Ranking Method on the System of Enterprise Accounts Annual Survey, Esprit SDC Project, Deliverable MI-3/D2.

[15] Domingo-Ferrer, J., Torra, V., (2001), A quantitative comparison of disclosure control methods for microdata, 111-133, in Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies, P. Doyle, J. I. Lane, J. J. M. Theeuwes, L. M. Zayatz (Eds.), Elsevier.

[16] Bacher, J., Brand, R., Bender, S., (2002), Re-identifying register data by survey data using cluster analysis: an empirical study, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10:5, 589-607.

[17] Torra, V., (2000), Re-identifying individuals using OWA operators, Proc. Int. Conf. Soft Comp., Iizuka, Japan.

[18] Yager, R. R., (1988), On ordered weighted averaging aggregation operators in multi-criteria decision making, IEEE Trans. on SMC, 18 183-190.

[19] Ohsawa, Y., (2001), The scope of chance discovery, Lecture Notes in Computer Science (LNAI subseries), 2253, 413.

[20] Murphy, P. M., Aha, D. W., (1994), UCI Repository Machine Learning Databases, http://www.ics.uci.edu/ mlearn/MLRepository.html, Irvine, CA: University of California, Department of Information and Computer Science.

[21] Domingo-Ferrer, J., Torra, V., Disclosure risk assessment in statistical microdata protection via advanced record linkage, Statistics and Computing, in press.