Record Linkage for Database Integration Using Fuzzy Integrals

Vicenç Torra,* Jordi Nin†
Institut d'Investigació en Intel·ligència Artificial (IIIA-CSIC), Campus UAB s/n, E-08193 Bellaterra, Catalonia, Spain

Given two-data databases, record linkage algorithms try to establish which records of these files contain information on the same individual. Standard record linkage algorithms assume that both files are described using the same attributes. In this article, we study the nonstandard case when the attributes are not the same. We apply aggregation operators for extracting relevant information for this purpose. We restrict to the case of numerical databases. © 2008 Wiley Periodicals, Inc.

1. INTRODUCTION

Re-identification algorithms are methods to establish relationships between data objects (records described in terms of attributes) that while supplied by different information sources correspond to the same entity. When sources correspond to databases and the data supplied correspond to records, two main classes of algorithms can be considered: methods for establishing correspondences between attributes[1] (schema matching algorithms) and methods for establishing correspondences between records of individuals[2,3] (record matching or record linkage).

Record linkage algorithms have been developed under different assumptions on the type of data available. Most algorithms assume that re-identification is applied to a pair of files (A, B) and that both files contain information on the same individuals. Some algorithms assume that both files contain exactly the same individuals[4] (so they try to find a one-to-one mapping between the two files). Also, algorithms usually assume that information is represented using numerical, categorical information, or textual attributes (*i.e.*, *strings*).

Standard record linkage further assumes that the two files A and B are described using the same attributes. Note that re-identification in this case is far from trivial because it is usually the case that files contain errors[5] (either accidental or on-purpose).

*Author to whom all correspondence should be addressed: e-mail: vtorra@iiia.csic.es.
†e-mail: jnin@iiia.csic.es.

INTERNATIONAL JOURNAL OF INTELLIGENT SYSTEMS, VOL. 23, 715–734 (2008)
 Published online in Wiley InterScience (www.interscience.wiley.com). • DOI 10.1002/int.20293

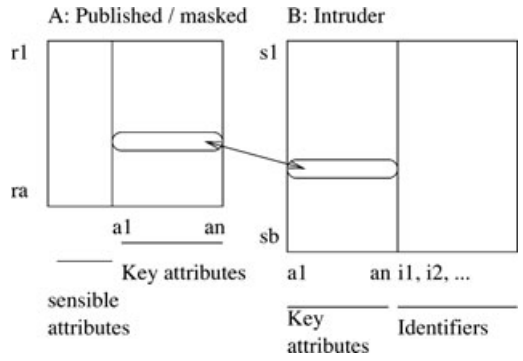

Figure 1. Record linkage for measuring disclosure risk.

1.1. Applications and Motivation

Record linkage techniques are used in several applications. To name a few, we have them in databases for data cleaning and in data mining for multidatabase data mining. Record linkage is a basic tool in privacy preserving data mining.[6,7] In this framework, it is used for assessing the disclosure risk before the publication of a file. The most common scenario is illustrated.

To ensure privacy of data files, files are protected using the so-called masking methods before publication. Such methods distort the original files in such a way that they are still valid for the intended data uses (*e.g.,* building models among attributes) and sensitive information cannot be inferred from the data. Thus, two elements need to be considered before publication: validity and privacy. Validity is usually evaluated in terms of information loss measures and the level of privacy can be evaluated using disclosure risk measures.[8–11]

Figure 1 illustrates the scenario where record linkage is used to measure disclosure risk. An intruder is assumed to have access to the published data (file A) and some personal data (file B), but that is not able to infer the relation between the records in the two files. Then, it is assumed that file B is, partly, described in terms of some key attributes. These key attributes are also used to describe the records in file A. Besides this, records in file B contain some additional information that corresponds to identifiers (*e.g.*, name, address, id-card number). Instead, records in file A also contain some additional information that corresponds to some sensitive attributes.

In this scenario, if the intruder correctly links a record a in A with the corresponding record b in B, the intruder would be able to disclose sensitive information contained in A for the individual who is identified using the identifiers in B.

Because of this, record linkage offers an appropriate tool to measure disclosure risk. Under this scenario, the protected data file A (the one to be published and available to all intruders) is compared with the original file B (the maximum amount of information that the intruder would have available). Then, the number of records in A that is correctly linked with its original records in B gives an estimation of the disclosure risk. For the sake of simplicity, we consider here that all attributes in A

(as well as all attributes in B) are key attributes. If this is not so, we would define A' and B' as files A and B restricted to key attributes. Then, we compare A' and B'.

Naturally, different record linkage algorithms give different estimations of disclosure risk. In this case, assuming that an intruder would use the best one, the risk corresponds to the largest number of re-identifications.

As explained above, standard record linkage methods assume that the records to be re-identified are described using the same attributes. That is, key attributes in A and B are the same. Nevertheless, other scenarios are also possible.

1.2. An Alternative Scenario

Torra[12] introduced a scenario in which files do not share the same attributes. In this case, re-identification is still possible if the attributes represent similar information. This would be the case, for example, if we have the attribute *Income-tax* in file B, whereas the file A contains *Net-income*.

In general, re-identification can still be achieved in this context under the following assumptions[12]:

Assumption 1. A set of common individuals is shared by both files.

Assumption 2. Data in both files contain, implicitly, similar structural information.

Following Torra,[12] we can say that "structural information of data is defined as any organization of the data that allows explicit representation of the relationship between individuals. This structural information is obtained from the data files through manipulation of the data (*e.g.*, using clustering techniques or any other data analysis or data mining technique)". In other words, although there are no common attributes, substantial correlation exists between some attributes in both files; or applying some clustering techniques, we obtain the same clusters for both sets of records.

Figure 2 represents a case that satisfies this latter assumption. In this case, two files A and B are considered. File A describes a set of retailers in terms of the attributes $\{$*Benefits, Start-up costs*$\}$ and file B describes the same retailers with the attribute *Business type*. In this case, some re-identifications are possible.

As different formalisms can be used for representing the structural information, different techniques are needed to extract such structural information. In this work, we focus on structural information represented by means of numerical representatives (as in the example of *Income-tax* vs. *Net-income*).

Assumption 3. Structural information is expressed by means of numerical representatives for each individual.

At present, there exist several works in the literature dealing with scenarios in which files do not include common attributes. Most of the research corresponds to attribute matching or schema matching.[1,13,14] Record linkage has been considered in Torra[12] and in Domingo-Ferrer and Torra[15] where the structural information

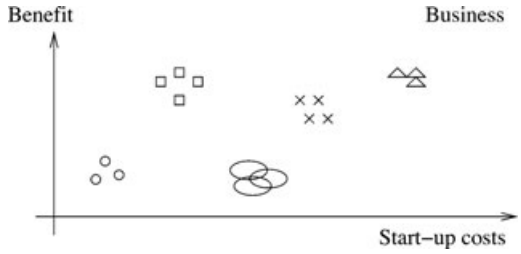

Figure 2. Graphical representation of an artificial problem that satisfies Assumption 2: File A with attributes {Benefits, Start-up costs} and File B with attribute {Business type}. In this figure, business types are represented using squares, ellipses, triangles, and so forth.

was described in terms of partitions (and extracted using clustering algorithms). Torra[16,17] considered the use of aggregation operators for re-identification. Nevertheless, the application described in such works corresponds, as pointed out by Malin and coworkers,[18,19] to attribute matching. Attribute matching is computationally simpler than record matching because the amount of redundant information existing in the data for attributes is larger than that for records.

In this article, we describe an approach for record linkage for the case that files do not share attributes and when the structural information is expressed using numerical representatives. We show that under a few conditions, aggregation operators arise as the appropriate functions for building such representatives. Aggregation operators, which are described in detail in Section 2.2, are functions that combine (aggregate) N values into a single one.

Aggregation operators have already been used in other information managing processes.[20–24]

The structure of this work is as follows. In Section 2, we describe some elements that are needed in the rest of the article. In Section 3, we describe our approach. In Section 4, we describe some experiments for its validation. The article finishes with some conclusions and directions for future work.

2. PRELIMINARIES

In this section, we review some concepts that are needed later on in this work. We start reviewing standard record linkage algorithms, and then we describe a few aggregation operators.

2.1. Record Linkage

Two main approaches exist for standard record linkage: probabilistic and distance-based record linkage. We detail them as follows:

- **Probabilistic record linkage**[5,25,26]**:** For each pair of records (a, b) where a is a record in file A and b is a record in B, we compute an index. Then, using such index, pairs

are classified as a linked pair (LP), a clerical pair (CP), or a nonlinked pair (NP). A clerical pair is one that cannot be automatically classified as linked or nonlinked (human inspection is needed to classify it).

In this framework, the indices are computed using conditional probabilities. Such probabilities are usually estimated using the expectation–maximization (EM) algorithm. Then, the thresholds are computed from (i) the probability of linking a pair that is an unmatched pair (a *false positive* or *false linkage*: $P(LP|\mathbf{U})$) and (ii) the probability of not linking a pair that is a match pair (a *false negative* or *false unlinkage*: $P(NP|\mathbf{M})$).

- **Distance-based record linkage**[27–29]**:** For each record a in A, we compute the distance to every record b in file B. Then, record a is linked to the nearest record in B. Distance between records is usually computed in terms of distance functions over attributes and assuming equal weight to all attributes. Nevertheless, ad hoc distances have been proven more successful in particular re-identification problems.[30]

In general, it can be said that distance-based record linkage is simpler to implement than probabilistic record linkage, and that permits to include easily subjective information through the distance function. However, the requirement of appropriate distance functions for each attribute and the establishment of weights (relative importance) between attributes makes the tuning phase a costly process. This is specially true both for categorical attributes on ordinal scales and for textual (*string*) attributes. Instead, in probabilistic record linkage, the parameters are easily learnt using the EM algorithm (*i.e.*, applying unsupervised methods). In this latter case, parameters are determined only from the probabilities of having both a false positive and a false negative linkages. Probabilistic record linkage is specially suited for categorical data because the conditional probabilities used correspond to coincidence (or noncoincidence) between pairs of values. Probabilities are prorated when partial coincidence is found. This is the case when numerical data are considered.

Both methods have been extensively tested[31] using as testbeds, sets of pairs of files with either numerical or categorical attributes. Each pair of files corresponded to one original file and a masked file obtained using a particular technique with a particular parameterizations. The results show that distance-based record linkage obtains a better performance when files contain numerical data and that probabilistic record linkage obtains a better performance when files contain categorical data.

In addition to the two methods described above, recently a third method was proposed.[32] Nevertheless, this latter method is similar to the distance-based records linkage.

2.2. Aggregation Operators

Aggregation operators[33] are numerical functions used for information fusion that combine N numerical values into a single one. These operators, which are formally described as follows, typically satisfy unanimity (idempotency) and monotonicity.

DEFINITION 1. *Let* $X := \{x_1, \ldots, x_N\}$ *be a set of information sources, and let* $f(x_i)$ *be a function to model that the* i*-th information source* x_i *supplies value* $f(x_i)$*,*

then a function $\mathbb{C} : \mathbb{R}^N \to \mathbb{R}$ *to aggregate values* $f(x_i)$ *is said to be an aggregation function if it satisfies:*

1. $\mathbb{C}(a, \ldots, a) = a$ *(unanimity, also known as idempotency)*
2. $\mathbb{C}(a_1, \ldots, a_N) \leq \mathbb{C}(a'_1, \ldots, a'_N)$ *if* $a_i < a'_i$ *(monotonicity)*

At present, several aggregation operators exist in the literature.[33,34] Among them, the most well-known aggregation operators are the arithmetic mean and the weighted mean. They correspond, respectively, to the following functions:

1. $\mathbb{C}(a_1, \ldots, a_N) = \frac{\sum_i^N a_i}{N}$
2. $\mathbb{C}(a_1, \ldots, a_N) = \sum_i^N w_i a_i$

In the second definition, $\mathbf{w} = (w_1 \ldots w_N)$ stands for a weighting vector. That is, w_i are weights for sources x_i such that $w_i \geq 0$ and $\sum_i w_i = 1$. These values correspond, using artificial intelligence jargon, to prior knowledge on the reliability of the sources. For example, when source x_i is twice as reliable as source x_j, then we have $w_i = 2w_j$.

Yager[35] defined the so-called ordered weighted averaging (OWA) operator that corresponds to a weighted linear combination of order statistics. At present, there are different definitions for this operator on the basis of the way the weights are defined. In this article, we recall a definition based on a nondecreasing function because this is the most useful definition in our context.

DEFINITION 2. *Let* Q *be a nondecreasing function in* $[0, 1]$ *such that* $Q(0) = 0$ *and* $Q(1) = 1$*, then the mapping* $OWA_Q : \mathbb{R}^N \to \mathbb{R}$ *defined as follows is an OWA operator:*

$$OWA_Q(a_1, \ldots, a_N) = \sum_{i=1}^{N} (Q(i/N) - Q((i-1)/N)) a_{\sigma(i)}$$

where σ *is a permutation of the values* a_i *such that* $a_{\sigma(i)} \geq a_{\sigma(i+1)}$.

This operator has several properties. We underline the following ones:

(i) For all Q, it holds that:

$$\min_i a_i \leq OWA_Q(a_1, \ldots, a_N) \leq \max_i a_i.$$

(ii) The function Q permits to modulate the output. For example, when we consider the family of functions $Q_\alpha(x) = x^\alpha$, we have that large positive values of α lead to an OWA near to the minimum and, instead, values of α near to zero lead to an OWA near to the maximum. Also, when a_i is fixed, OWA_{Q_α} is nondecreasing with respect to α. These conditions are formalized as:

- $\lim_{\alpha\to\infty} OWA_{Q_\alpha}(a_1, \ldots, a_N) = a_{\alpha(N)} = \min a_i$
- $\lim_{\alpha\to 0} OWA_{Q_\alpha}(a_1, \ldots, a_N) = a_{\alpha(1)} = \max a_i$
- if $\alpha_1 > \alpha_2$, then $OWA_{\alpha_1}(a_1, \ldots, a_N) < OWA_{\alpha_2}(a_1, \ldots, a_N)$

(iii) The OWA operator is symmetric for all Q. That is, the order of the parameters is not relevant for the computation of the output. This can be formalized as follows:

$$OWA_Q(a_1, \ldots, a_N) = OWA_Q(a_{\pi(1)}, \ldots, a_{\pi(N)})$$

for any permutation π.

Another relevant property of OWA operators is that they are equivalent to the so-called Choquet integrals[36] with respect to symmetric fuzzy measures. Choquet integrals are one family of the so-called fuzzy integrals,[37] a set of functionals that can be used for information fusion. In short, given a function f that represents the information supplied by the sources in X, the fuzzy integral of f with respect to a fuzzy measure represents an aggregated value of those values in f. In such integrals, fuzzy measures play the role of weights in the weighted mean (*i.e.*, some prior knowledge on the reliability of the sources). The main difference between a fuzzy integral and a weighted mean is that in the weighted mean independence is assumed among the information sources. Instead, such independence is not formally required for fuzzy integrals because fuzzy measures can accommodate the dependencies among the sources.

Formally speaking, a fuzzy measure is a set function over X ($\mu : 2^X \to [0, 1]$) that satisfies the following constraints:

- $\mu(\emptyset) = 0$, $\mu(X) = 1$ (boundary conditions).
- if $A \subseteq B$ then $\mu(A) \leq \mu(B)$ (monotonicity conditions).

The OWA operator of f with respect to Q is equivalent to the Choquet integral of f with respect to the fuzzy measure μ defined as: $\mu(A) = Q(|A|/N)$, where $|\cdot|$ stands for the cardinality of a set. This equivalence establishes that the fuzzy measure associated to the OWA for a set A does not depend on the particular elements in A but only on its cardinality. That is, given two sets $A \neq B$ ($A, B \subseteq X$) such that $|A| = |B|$, then $\mu(A) = \mu(B)$. Because of this, the measure is said to be symmetric and, naturally, any Choquet integral with respect to a measure of this form is also symmetric because this corresponds to the OWA operator.

The property that a Choquet integral with respect to a symmetric fuzzy measure is symmetric also holds for other fuzzy integrals. In particular, it also holds for the Sugeno integral.[38] We give the definition of the Sugeno integral with respect to a symmetric fuzzy measure representable, as above, in terms of a function Q. This expression is equivalent to the OWMax defined by Yager[39] (that is related with the ordinal WMax,[40] but with an ordering of the data).

DEFINITION 3. *Let Q be a nondecreasing function in* $[0, 1]$ *such that* $Q(0) = 0$ *and* $Q(1) = 1$, *then the mapping* $SI_Q : \mathbb{R}^N \to \mathbb{R}$ *defined as follows is a Sugeno integral with respect to the fuzzy measure* $\mu(A) = Q(|A|/N)$*:*

$$SI_Q(a_1, \ldots, a_N) = \max_{i=1}^{N} min(Q(i/N), a_{\sigma(i)})$$

where σ *is a permutation such that* $a_{\sigma(i)} \geq a_{\sigma(i+1)}$.

As stated above, this function is symmetric for all Q. Besides this, the function is an aggregation operator (in the sense of Definition 1) and the function Q permits to modulate the output of the integral.

See Torra and Narukawa[33] for additional details on aggregation operators and fuzzy integrals.

3. PROCEDURE FOR RE-IDENTIFICATION

As we have explained in the introduction, we consider here the re-identification of records between two files A and B when no shared attributes are present between the two files. Accordingly, the methods described in Section 2.1 cannot be applied.

To tackle this problem, we consider the transformation of files A and B into two new files A' and B' with the goal that standard re-identification algorithms can be applied on this latter pair of files (A', B').

To do so, we consider the construction of several representatives for each record a in A and each record b in B so that re-identification can be performed over such representatives. This process is detailed below:

- First, we consider a set of functions f_i for building the representatives. In general, we consider that f_i is a function of both the record and the whole data file A. Therefore, being a a record in A, $f_i(a, A)$ stands for a representative of the record. We denote the set of considered functions by $\mathcal{F} = \{f_i\}$ for $i = 1, \ldots, k$.
- Then, we apply the functions in $\mathcal{F}$ to the records a in A to obtain a'. Formally speaking, $a' := \mathcal{F}(a, A)$ where:

$$a' := \mathcal{F}(a, A) = (f_1(a, A), \ldots, f_k(a, A))$$

- Now, assuming that functions in $\mathcal{F}$ are also applicable to records b in B, we define records b' in B in a similar way:

$$b' := \mathcal{F}(b, B) = (f_1(b, B), \ldots, f_k(b, B))$$

- Finally, we define files A' and B' in terms of the new records a' and b'. That is:

$$A' := \{\mathcal{F}(a, A)\}_{a \in A}$$

$$B' := \{\mathcal{F}(b, B)\}_{b \in B}$$

Therefore, given the set of functions $\mathcal{F} = \{f_i\}$ for $i = 1, \ldots, k$, and applying each f_i to each record in A and B, we obtain files A' and B'. In a procedural way, this process is defined as follows:

```
1 function transformFile (A: File, F: set of
  functions) is
2 A':File; // Records in file A' will contain |F|
  numerical attributes;
3 for each record a in A
4    a':=new record(f1(a, A), ..., fk(a, A));
5    write(a',A');
6 end for;
7 return A';
8 end function;
```

Therefore, files A' and B' will be obtained as:

```
A':=transformFile(A, F); B':=transformFile(B, F);
```

With this construction, both files A' and B' contain the same number of records as A and B and records in both files are described using the same kind of representatives. Therefore, both files can be considered as described using the same attributes and, as such, standard re-identification algorithms can be applied to the pair (A', B').

At this point, it is clear that a crucial decision is the selection of functions in $\mathcal{F}$. This is considered in detail in the next section.

3.1. Aggregation Operators for Building Representatives

For building the representatives, we have to select the functions in $\mathcal{F}$. First, we show that aggregation operators are suitable functions for this purpose, and then, on the basis of the properties we require for f_i, we show that the OWA is an appropriate selection.

So, we turn into the requirements for functions $f \in \mathcal{F}$:

(i) The outcome of f applied to a record a should not depend on the values of the other records in A. This condition corresponds to the so-called *condition of independence of irrelevant alternatives*, and its inclusion excludes functions based, for example, on principal component analysis. Formally speaking, this condition implies that functions $f(a, A)$ do depend only on a and should not depend on the other values in A.

(ii) When all the values of a record are equal, this very value is the representative. This condition implies that all functions f satisfy unanimity (idempotency).

(iii) The representatives should be monotonic with respect to its inputs. That is, given two records $a = (a_1, \ldots, a_N)$ and $a^* = (a_1^*, \ldots, a_N^*)$ such that $a_i \leq a_i^*$, the representatives of a should always be smaller than (or equal to) the representatives of a^*.

(iv) When there is no prior knowledge on the attributes (if this is not the case, other methods might be used for linkage), no preference should be given to any of the attributes involved in the process. In other words, the order of the attributes is irrelevant. This is formally expressed saying that a permutation of the attributes does not affect the output:

$$f(a_1, \ldots, a_N) = f(a_{\pi(1)}, \ldots, a_{\pi(N)}),$$

where π is a permutation of the indices. That is, f is a symmetric function.

(v) The function should be easily extensible to an arbitrary number of parameters, so that the same procedure can be applied to files with an arbitrary number of attributes. In this

way, we can apply $\mathcal{F}$ to both files A and B, although the number of attributes in each file is different.

(vi) The function should be parameterizable so that different representatives can be computed for the same record.

These requirements constraint functions in $\mathcal{F}$. In particular, the first condition implies that functions $f_i(a, A)$ can be defined in terms of another function f_i' that depends only on a. That is, $f_i(a, A) = f_i'(a)$. Then, conditions (ii) and (iii) imply that functions f_i' are aggregation operators because they should be idempotent and monotonic (see Definition 1). Therefore, the following holds:

PROPOSITION 1. *Let the functions in $\mathcal{F}$ satisfy the condition of independence of irrelevant alternatives, idempotency, and monotonicity. Then, the functions in $\mathcal{F}$ are aggregation operators.*

In addition, when conditions (iv), (v), and (vi) are required for aggregation operators, some of such operators are discarded. This is the case of the weighted mean (that is not symmetric and not easily extensible because it requires weights for each attribute) or the arithmetic mean (that is not parameterizable). The OWA operator and the other fuzzy integrals with symmetric fuzzy measures are some of the few ones that are appropriate. They are symmetric and parameterizable (in terms of the function Q). In relation to the property of being extensible for an arbitrary number of parameters, we have that not all definitions for OWA operators are appropriate. For example, definitions based on weighting vectors (as the original definition in Yager[35]) are not appropriate because additional arguments would require additional weights. Nevertheless, the definition given in Definition 2 is appropriate because the same function Q can be used for an arbitrary value of N.

Taking all this into account, we can use either OWA operators or Sugeno integrals (both based on nondecreasing functions Q). This selection is valid as the following proposition holds:

PROPOSITION 2. *The functions OWA_Q and SI_Q satisfy conditions (i)–(vi) for all nondecreasing functions Q.*

In addition, as functions satisfying condition (v) are applicable to an arbitrary number of parameters, they can also be applied to situations in which data contain missing values. In this case, instead of defining record a' as before, we would define:

$$a' := \mathcal{F}(a, A) = (f_1(\hat{a}, A), \ldots, f_k(\hat{a}, A)),$$

where $\hat{a}$ is a projection of a over those attributes with nonmissing values in a. For all this, the following holds:

PROPOSITION 3. *The functions OWA_Q and SI_Q are applicable to records with missing data, for all nondecreasing functions Q.*

3.2. Example

Now, we illustrate the method we have proposed with a simple example. In Section 4, we will describe several experiments with real data in detail.

Let us consider the two data files A and B represented in Tables I and II. File A consists of 10 records described in terms of four attributes. All attributes are numerical and numbers belong to the interval [0, 1]. File B contains the same data included in file A, but the attributes have been permuted.

Standard re-identification algorithms cannot be applied to establish links between the records in A and B without knowing the correspondence between attributes in A and B. Nevertheless, in this case, we can apply the method described in this section. To do so, we need to define the set of functions $\mathcal{F}$. We use here the OWA operator using $Q_\alpha(x) = x^\alpha$ with several values of α. In particular, we consider 10 different functions corresponding to Q_α with the following values of α:

$$\alpha = (1/5, 2/5, 3/5, 4/5, 5/5, 6/5, 7/5, 8/5, 9/5, 10/5).$$

Applying these aggregation functions, we obtain exactly the same records for both files A and B in Tables I and II. The records obtained are given in Table III. Now, as both files contain exactly the same records, the re-identification is trivial.

Table I. File A for re-identification.

Record	V_1	V_2	V_3	V_4
r_1^A	0.2	0.4	0.2	0.4
r_2^A	0.1	0.2	0.1	0.2
r_3^A	0.5	0.6	0.5	0.1
r_4^A	0.8	0.4	0.4	0.7
r_5^A	0.9	0.2	0	0
r_6^A	0.2	0.2	0.3	0.9
r_7^A	0.5	0.3	0.2	1
r_8^A	0	0.1	0.5	1
r_9^A	1	0	0.9	0.2
r_{10}^A	0.5	1	1	0.8

Table II. File B for re-identification.

Record	V_1'	V_2'	V_3'	V_4'
r_1^B	0.4	0.2	0.2	0.4
r_2^B	0.2	0.1	0.1	0.2
r_3^B	0.1	0.5	0.5	0.6
r_4^B	0.7	0.4	0.8	0.4
r_5^B	0	0	0.9	0.2
r_6^B	0.9	0.3	0.2	0.2
r_7^B	1	0.2	0.5	0.3
r_8^B	1	0.5	0	0.1
r_9^B	0.2	0.9	1	0
r_{10}^B	0.8	1	0.5	1

Table III. File A (and B) for re-identification.

Record	$Q_{1/5}$	$Q_{2/5}$	$Q_{3/5}$	$Q_{4/5}$	$Q_{5/5}$	$Q_{6/5}$	$Q_{7/5}$	$Q_{8/5}$	$Q_{9/5}$	$Q_{10/5}$
$r'^A_1 = r'^B_1$	0.374	0.351	0.332	0.315	0.3	0.287	0.276	0.266	0.257	0.25
$r'^A_2 = r'^B_2$	0.187	0.176	0.166	0.157	0.15	0.144	0.138	0.133	0.129	0.125
$r'^A_3 = r'^B_3$	0.553	0.514	0.480	0.451	0.425	0.402	0.382	0.363	0.347	0.33125
$r'^A_4 = r'^B_4$	0.737	0.685	0.641	0.605	0.575	0.550	0.528	0.510	0.494	0.48125
$r'^A_5 = r'^B_5$	0.705	0.554	0.437	0.346	0.275	0.220	0.176	0.142	0.115	0.09375
$r'^A_6 = r'^B_6$	0.742	0.620	0.527	0.455	0.4	0.357	0.324	0.298	0.278	0.2625
$r'^A_7 = r'^B_7$	0.847	0.728	0.634	0.559	0.5	0.453	0.414	0.383	0.358	0.3375
$r'^A_8 = r'^B_8$	0.822	0.679	0.566	0.474	0.4	0.340	0.290	0.249	0.216	0.1875
$r'^A_9 = r'^B_9$	0.874	0.766	0.674	0.594	0.525	0.465	0.413	0.368	0.328	0.29375
$r'^A_{10} = r'^B_{10}$	0.957	0.919	0.884	0.853	0.825	0.799	0.776	0.755	0.736	0.71875

Note that the first row in Table III is obtained applying the OWA operator to the first row of Table I (and Table II) using the function $Q_\alpha(x) = x^\alpha$ with $\alpha = 1/5, \ldots, 10/5$. In particular, the first column in Table III corresponds to $\alpha = 1/5$, second column to $\alpha = 2/5$, and so forth until the tenth column where $\alpha = 10/5$.

Therefore, the element in the i-th row, column Q_α in Table III, corresponds to $OWA_{Q_\alpha}(r_i^A)$. Of course, $OWA_{Q_\alpha}(r_i^A)$ is equivalent to $OWA_{Q_\alpha}(r_i^B)$ in this example because r_i^B is a permutation of r_i^A and the OWA operator is symmetric.

This example can be considered as too simplistic. Nevertheless, this same situation arises in database integration with unlabeled attributes or with inconsistent labeled attributes. In a more general case, instead of having a permutation of exactly the same attributes, we might have attributes in one file that are combination of some attributes in the other database.

4. EXPERIMENTS

The approach presented here has been extensively tested with several data files, considering both types of aggregation operators (OWA and simplified Sugeno integral) and considering three different quantifiers. In this section, we review the experiments performed.

4.1. Data Files

We have considered seven files from the UCI repository[41] and one file extracted using the Data Extraction System (DES) from the U.S. Census Bureau.[42] The names of the files from the UCI repository are listed in the next section. The file from the U.S. Census Bureau contains data from the "Current Population Survey" (corresponding to 1995) and, more specifically, to the file group "March Questionnaire Supplement—Person Data Files." We refer to this file as *census* as it was extracted from the U.S. Census Bureau DES.

As the method described in this work is defined only for numerical data, we have selected files described in terms of numerical attributes. Nonnumerical attributes, if any, were discarded.

4.2. Preprocessing

The files have been partitioned to test the re-identification algorithms. Each file was split into two new files in such a way that both files contained the same records but only some of the attributes. Attribute selection was done on the basis of the correlation coefficients. In particular, attributes with a low correlation coefficient with all the other attributes were discarded and pairs of attributes with a correlation coefficient of at least 0.7 were separated, putting one in each file.

We list the databases considered, and for each file two sets of attributes are considered (each set defines one file). For example, in the case of the Iris Plants Database, which contains 150 records, one file contains 150 records but only attributes *Sepal-length* and *Petal-length* and the other file (which also contains 150 records) contains attributes *Sepal-width* and *Petal-width*.

- Iris Plants Database: {sepal length, petal length}, {sepal width, petal width}
- Abalone Database: {Height, Whole weight, Viscera weight}, {Length, Diameter, Shucked weight, Shell weight}
- Ionosphere Database: $\{V_5, V_7, V_9, V_{11}, V_{13}, V_{20}\}$, $\{V_{15}, V_{17}, V_{19}, V_{21}, V_{23}, V_{30}\}$
- Dermatology Database: {polygonal papules, follicular papules, oral mucosal involvement, knee and elbow involvement, scalp involvement, melanin incontinence, exocytosis, focal hypergranulosis, follicular horn plug}, {clubbing of the rete ridges, elongation of the rete ridges, thinning of the suprapapillary epidermis, vacuolisation and damage of basal layer, saw-tooth appearance of retes, perifollicular parakeratosis, band-like infiltrate}
- Boston Housing Data: {INDUS, RM, AGE, RAD}, {NOX, TAX, MEDV}
- Faults in an urban waste water treatment plant (Water-treatment): {PH-E, DBO-E, SS-E, SSV-E, SED-E, COND-E, DBO-D, SSV-D, DBO-S, RD-DBO-S, RD-DQO-S}, {PH-P, DBO-P, SS-P, SSV-P, SED-P, COND-P, PH-D, DQO-D, COND-D, SS-S, SED-S, COND-S, RD-DBO-G, RD-DQO-G}
- Wisconsin Diagnostic Breast Cancer (WDBC): $\{V_2 - V_4, V_6 - V_8, V_{10}, V_{12}, V_{13}, V_{18}, V_{20}, V_{26}, V_{29}, V_{32}\}$, $\{V_5, V_9, V_{15} - V_{16}, V_{19}, V_{22} - V_{25}, V_{27} - V_{28}, V_{30}\}$
- 1995 - Current Population Survey (Census): {AFNLWGT, EMCONTRB, PTOTVAL, TAXINC, POTHVAL, PEARNVAL, WSALVAL}, {AGI, FEDTAX, STATETAX, INTVAL, FICA, ERNVAL}

Before applying the re-identification algorithm, the data have been normalized. We have considered both ranging (denoted by N_1) and standardization (N_2). Missing values have been replaced by zero (after normalization).

4.3. Tests

The procedure described in Section 3 has been applied to each pair of files. For each pair, we have selected at random sets of 100 records and applied the re-identification algorithms. Ten executions have been applied and the average number of re-identifications has been computed.

Experiments have been done for both the OWA operator and the Sugeno integral with respect to a fuzzy measure of the form $\mu(A) = Q(|A|/N)$. For both operators, three different families of nondecreasing functions were considered. The functions

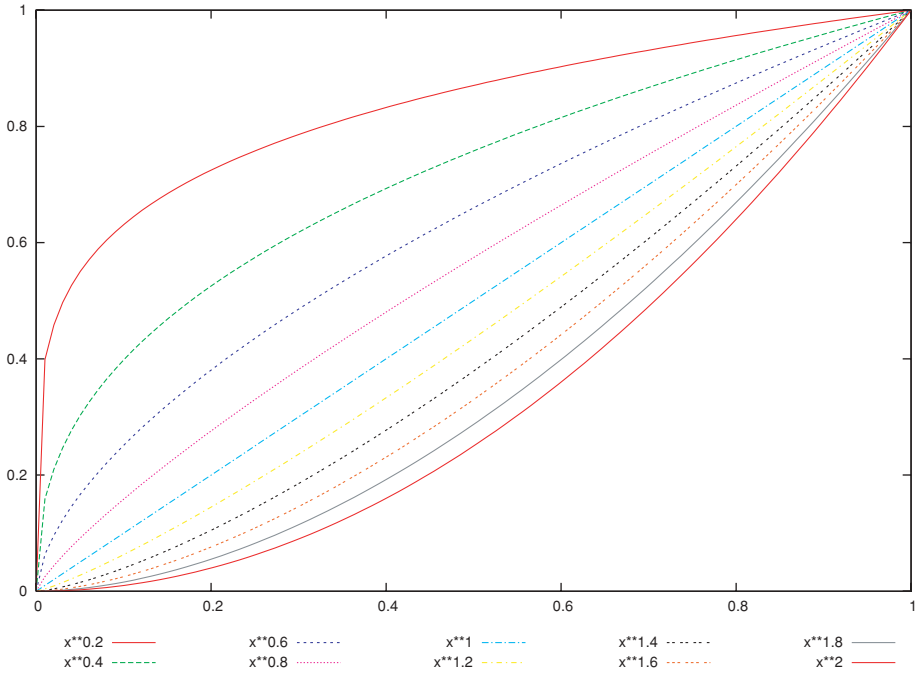

Figure 3. Graphical representation of Q^e_α for $\alpha = 1/5, \ldots, 10/5$.

and the parameters used are as follows:

(1) $Q^e_\alpha(x) = x^\alpha$ for $\alpha = 1/5, 2/5, 3/5, \ldots, 10/5$
(2) $Q^s_\alpha(x) = 1/(1 + e^{(\alpha - x)*10})$ for $\alpha = \{0, 0.1, \ldots 0.9\}$
(3) $Q^t_\alpha(x) = \begin{cases} 0 \text{ if } x \leq \alpha \\ 1 \text{ if } x > \alpha \end{cases}$ for $\alpha = \{0, 0.1, \ldots 0.9\}$

Here, Q^e stands for exponent, Q^s for sigmoidal, and Q^t for threshold. Figures 3–5 give a graphical representation of these functions.

Once we have files with common attributes, both probabilistic and distance-based record linkage have been used.

4.4. Results

The results for the files with better performance are given in Tables V–VIII. They correspond to the files abalone, ionosphere, census, and wdbc. Iris, Dermatology, and housing lead to poor results. The bad performance of iris and housing was probably due to the small number of attributes which did not permit to express structural information correctly. The file water-treatment, not included here, led to results similar to census, with around 10 re-identifications and a maximum of 16.

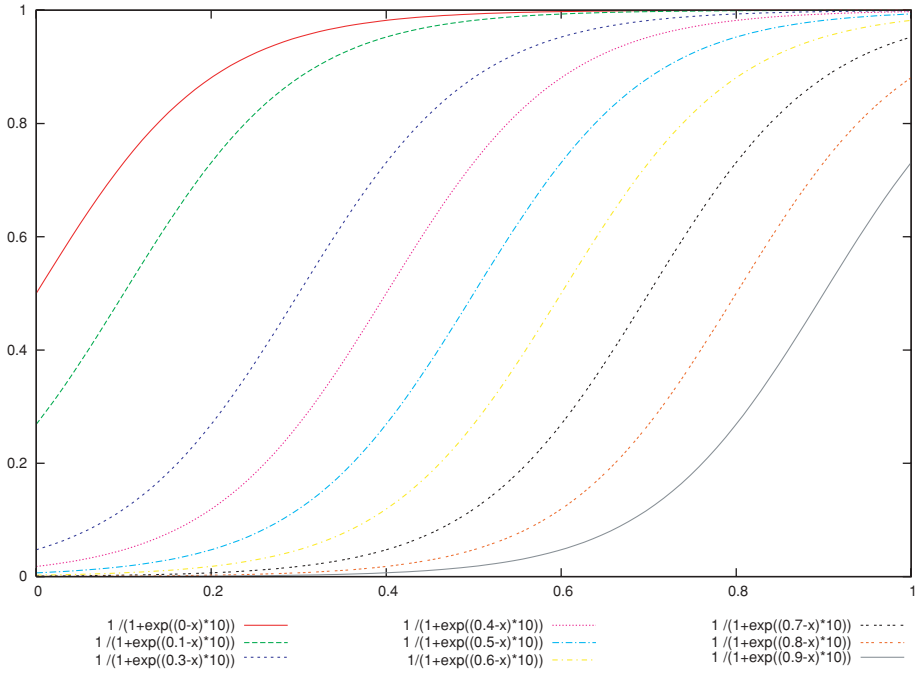

Figure 4. Graphical representation of Q_α^s for $\alpha = 0, 0.1, \ldots, 0.9$.

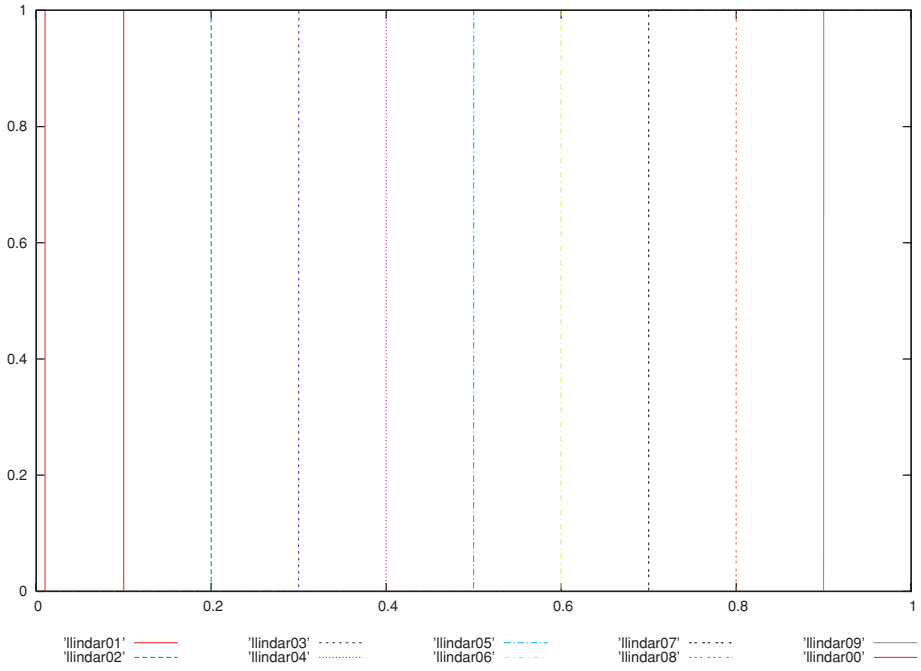

Figure 5. Graphical representation of Q_α^t for $\alpha = 0, 0.1, \ldots, 0.9$.

Table IV. Probabilities of having both r correct links and equal or more than r links for 100 records.

r	prob. $\|links\| = r$	prob. $\|links\| \geq r$
0	0.36787944	1
1	0.36787944	0.63212056
2	0.18393972	0.26424112
3	0.06131324	0.08030140
5	0.00306566	0.00365985
10	1.0138E-7	1.1143E-7
15	2.8132E-13	3.0000E-13
20	1.5121E-19	1.5875E-19
26	9.1219E-28	9.4723E-28
28	1.2066E-30	1.2496E-30
30	1.3869E-33	1.4331E-33
50	1.2096E-65	1.2338E-65
100	1.071E-158	1.071E-158

Table V. Average number of re-identified records for the Abalone example.

		OWA			SI		
		Q^e	Q^s	Q^t	Q^e	Q^s	Q^t
N_1	DBRL	6.5	5.9	6.7	4.8	4.2	6.7
	PRL	3.9	5.2	1.8	5.5	5.2	1.8
N_2	DBRL	9.9	7.9	8.8	5.6	6.5	7.0
	PRL	6.3	8.4	2.2	5.6	6.2	2.4

Table VI. Average number of re-identified records for the Ionosphere example.

		OWA			SI		
		Q^e	Q^s	Q^t	Q^e	Q^s	Q^t
N_1	DBRL	14.4	21.8	21.9	11.6	20.3	21.9
	PRL	12.9	22.2	3.9	10.8	20.7	3.9
N_2	DBRL	5.7	7.9	8.6	6.4	6.9	8.0
	PRL	4.2	7.5	1.3	4.9	6.2	1.6

Table VII. Average number of re-identified records for the Census example.

		OWA			SI		
		Q^e	Q^s	Q^t	Q^e	Q^s	Q^t
N_1	DBRL	7.1	9.5	7.5	6.1	8.6	7.5
	PRL	4.7	9.6	10.4	6.0	7.9	10.4
N_2	DBRL	8.4	8.8	9.9	4.3	3.6	5.0
	PRL	7.4	8.8	5.0	3.7	3.5	2.2

Table VIII. Average number of re-identified records for the WDBC example.

		OWA			SI		
		Q^e	Q^s	Q^t	Q^e	Q^s	Q^t
N_1	DBRL	5.0	7.0	4.4	5.5	5.8	4.4
	PRL	4.4	7.1	8.0	6.3	5.8	8.0
N_2	DBRL	10.8	15.8	18.2	3.3	4.6	5.1
	PRL	10.5	14.8	16.2	3.3	4.7	4.6

In the tables,we give the average number of re-identifications obtained over 10 executions, considering in each execution the parameters described above: (i) either the OWA operator or the Sugeno integral with respect to a symmetric fuzzy measure (denoted as SI); (ii) either distance-based record linkage (DBRL) or probabilistic one (PRL); (iii) either ranging (denoted N_1) or standardization (N_2) as the normalization method; and (iv) Q^e, Q^s, or Q^t as the nondecreasing functions Q that define the set $\mathcal{F}$ with OWA or SI.

The experiments show that, except for the files with poor performance, at least 10% of the records were re-identified achieving averages of 18 and 21.9 for files wdbc and ionosphere. The maximum number of records re-identified in an experiment was 26 in wdbc and 28 in ionosphere. These values are not given in the tables, as such tables give only the averages of 10 executions.

The evaluation of our approach is not straightforward because there are no systematic alternative approaches to deal with the same problem. Two simple methods were considered[15] for the census problem:

- The one-dimensional ranking based on first principal component: this permitted to correctly re-identify 5 of 90 records.
- The one-dimensional ranking based on the sum of z-scores: Using this approach, 5 of 90 records were correctly re-identified.

For the same problem, using the approach described here, we were able to correctly re-identify 12 records of 100 and the better average over 10 runs is 10.4 (see Table VII).

An alternative way to assess the successfulness of the method is to consider the probability of random linkage. The probability of randomly obtaining r or more linkage out of n is defined in the next proposition.

PROPOSITION 4. [17,43] *If A and B both contain n records corresponding to the same set of n individuals, the probability of correctly re-identifying exactly r individuals by a random strategy is*

$$\frac{\sum_{v=0}^{n-r} \frac{(-1)^v}{v!}}{r!} \tag{1}$$

Table IV gives the probabilities for some values of r when the number of records is 100. It can be seen that the probability of obtaining between 15 and 30 records

(as obtained in some of the experiments reported here) is almost zero. For example, the probability of re-identifying 26 records or more as in wdbc is 9.47×10^{-28} and for 28 records as in ionosphere is 1.24×10^{-30}.

Finally, it is possible to compare the results of our approach with the success rate of re-identification of standard record linkage when an original file and a masked file are compared. Domingo-Ferrer and Torra[31] described around 300 experiments and an average number of re-identifications of 26.12% was obtained for distance-based record linkage and 19.72% for probabilistic one. Here, the rate is smaller, but considering that we use files not sharing attributes the performance is acceptable, specially as the best performance for ionosphere is 28 (>26.12%) and the best average for the same problem for probabilistic record linkage is 22.2%, still larger than the result of Domingo-Ferrer and Torra[31] for probabilistic record linkage. Similar results were reported[44] with respect to re-identification of synthetic data.

The results permit to compare the different approaches experimented. In general, we can say that the use of the Choquet integral is more successful than that of the Sugeno integral. Also, that the use of the quantifier Q^t leads to better results than the use of Q^e and Q^s. The results also show that distance-based record linkage is more suitable for numerical data. Finally, the use of standardization is, in general, preferable over ranging.

5. CONCLUSIONS AND FUTURE WORK

In this article, we studied a method for record linkage when files do not share attributes. Exhaustive testing has been carried out to evaluate its performance. Results show that the method permits an average re-identification of 10% to 20% of the records, achieving better performance for some of the files. As this results are significant, our approach is validated.

The experiments reported here consider files with only a limited number of records (100 records). Further work is needed to study the validity of the approach with larger files. Nevertheless, when larger files are considered, it is a common practice to split the problem into smaller problems using the so-called blocking technique. In this case, a combination of (categorical) attributes are used to partition the file, and then comparison between records is limited to records with the same values with such blocking attributes. In this way, the number of records to be compared is largely reduced and kept into reasonable size (at most a few thousands of records).

In the context of privacy, instead of matching two files with the same number of records (as done here), one of the files would have less records. In particular, one of the intruder will typically match on a few records (a subset of the published ones). In this case, the same procedure applies. The results obtained here give an estimate of the number of records that an intruder might re-identify.

Acknowledgments

Partial support by the Spanish MEC (projects ARES – CONSOLIDER INGENIO 2010 CSD2007-00004 – and eAEGIS – TSI2007-65406-C03-02) is acknowledged.

Jordi Nin thanks the Spanish Council for Scientific Research (CSIC) for his I3P grant.

References

1. Do H-H, Rahm E. COMA—A system for flexible combination of schema matching approaches. In: Proc 28th VLDB Conference, Hong-Kong, China; 2002.
2. Winkler WE. Data cleaning methods. In: Proc SIGKDD 2003, Washington, DC; 2003.
3. Winkler WE. Re-identification methods for masked microdata. In: Privacy in Statistical Databases 2004. Lecture Notes in Computer Science, vol 3050. Berlin: Springer; 2004. pp 216–230.
4. Robinson-Cox JF. A record-linkage approach to imputation of missing data: analyzing tag retention in a tag-recapture experiment. J Agric Biol Environ Stat 1998;3:48–61.
5. Winkler WE. Matching and record linkage. In: Cox BG, editor. Business survey methods. New York: Wiley; 1995. pp 355–384.
6. Agrawal R, Srikant R. Privacy preserving data mining. In: Proc the ACM SIGMOD Conf on Management of Data; 2000. pp 439-450.
7. Verykios VS, Bertino E, Nai Fovino I, Parasiliti L, Saygin Y, Theodoridis Y. State-of-the-art in privacy preserving data mining, SIGMOD Record 2004;33(1):50–57.
8. Agrawal D, Aggarwal CC. On the design and quantification of privacy preserving data mining algorithms. In Proc 20th ACM Symposium on Principle of Database System; 2001. pp 247–255.
9. Bertino E, Fovino IN, Provenza LP. A framework for evaluating privacy preserving data mining algorithms. Data Mining Knowl Discov 2005;11(2):121–154.
10. Domingo-Ferrer J, Torra V. Disclosure control methods and information loss for microdata, In: Doyle P, Lane JI, Theeuwes JJM, Zayatz LM. editors. Confidentiality, disclosure, and data access: theory and practical applications for statistical agencies. Amsterdam: Elsevier; 2001. pp 91–110.
11. Domingo-Ferrer J, Torra V. A quantitative comparison of disclosure control methods for microdata. In: Doyle P, Lane JI, Theeuwes JJM, Zayatz LM. editors. Confidentiality, disclosure, and data access: theory and practical applications for statistical agencies. Amsterdam: Elsevier; 2001. pp 111–133.
12. Torra V. Towards the re-identification of individuals in data files with non-common variables. In: Proc 14th European Conf on Artificial Intelligence (ECAI2000) (ISBN 1 58603 013 2). Berlin, Germany. IOS Press; 2000. pp 326–330.
13. Rahm E, Hai Do H. Data Cleaning: Problems and Current Approaches, Bull Tech Com Data Eng 2000;23(4):3–13.
14. Rahm E, Bernstein PA. A survey of approaches to automatic schema matching, VLDB J 2001;10:334–350.
15. Domingo-Ferrer J, Torra V. Disclosure risk assessment in statistical microdata protection via advanced record linkage. Stat Comput 2003;13:343–354.
16. Torra V. Re-identifying Individuals using OWA operators. In: Proc 6th Int Conf Soft Computing (Iizuka 2000) (ISBN 4-938717-05-0), (full paper in the electronic proceedings - CD Rom), Iizuka, Fukuoka, Japan; 2000. p 101.
17. Torra V. OWA operators in data modeling and reidentification. IEEE Trans Fuzzy Syst 2004;12(5):652–660.
18. Malin B, Sweeney L, Newton E. Trail Re-identification: Learning who you are from where you have been. In: LIDAP-WP12, Carnegie Mellon University, Lab. for Int'l Data Privacy, Pittsburgh, PA; March 2003.
19. Malin B. Betrayed by my shadow: learning data identity via trail matching, J Privacy Technol. 2005. Available at: http://www.jopt.org.
20. Bordogna G, Pasi G. Linguistic aggregation operators of selection criteria in fuzzy information retrieval. Int J Intel Syst 1995;10:233–248.

21. Herrera-Viedma E, Pasi G. Soft Approaches to Information Retrieval and Information Access on the Web. J Am Soc Inf Sci Technol 2006;57(4):511–514.
22. Herrera-Viedma E, Peis E. Evaluating the informative quality of documents in SGML format from judgements by means of fuzzy linguistic techniques based on computing with words, Inf Process Manage 2003;39:233–249.
23. Pasi G. Flexible Information Retrieval: some research trends. Mathware Soft Comput 2002;9(1):107–121.
24. Yager RR. Data mining using granular linguistic summaries. In: Torra V, editor. Information fusion in data mining. New York: Springer; 2003. pp 211–229.
25. Fellegi IP, Sunter AB. A theory for record linkage. J Am Stat Assoc 1969;64:328,1183–1210.
26. Jaro MA. Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida, J Am Stat Assoc 1989;84:406,414–420.
27. Tendick P. Assessing the effectiveness of the noise addition method of preserving confidentiality in the multivariate normal case. J Stat Plann Inference 1992;31:273–282.
28. Fuller WA. Masking procedures for microdata disclosure limitation. J Off Stat 1993;9:383–406.
29. Pagliuca D, Seri G. Some results of individual ranking method on the system of enterprise accounts annual survey, Esprit SDC Project, Deliverable MI-3/D2. 1999.
30. Torra V, Miyamoto S. Evaluating fuzzy clustering algorithms for microdata protection. Lecture Notes in Computer Science, vol 3050. Berlin: Springer; 2004. pp 175–186.
31. Domingo-Ferrer J, Torra V. Validating distance-based record linkage with probabilistic record linkage. In: Escrig MT, Toledo F. Golobardes E. editors. Topics in artificial intelligence. Lecture Notes in Artificial Intelligence. vol 2504. Berlin: Springer; 2002. pp 207–215.
32. Bacher J, Brand R, Bender S. Re-identifying register data by survey data using cluster analysis: an empirical study. Int J Uncertainty Fuzziness Knowl-Based Syst 2002;10(5):589–608.
33. Torra V, Narukawa Y. Modeling decisions: information fusion and aggregation operators. New York: Springer; 2007.
34. Calvo T, Mayor G, Mesiar R. Aggregation operators. Heidelberg: Physica-Verlag; 2002.
35. Yager RR. On ordered weighted averaging aggregation operators in multi-criteria decision making. IEEE Trans Syst Man Cybern 1988;18:183–190.
36. Choquet G. Theory of capacities. Ann Inst Fourier 1954;5:131–296.
37. Grabisch M, Murofushi T, Sugeno M. Fuzzy measures and integrals: theory and applications. Heidelberg: Physica-Verlag; 2000.
38. Sugeno M. Theory of fuzzy integrals and its application. Thesis, Tokyo Institute of Technology, 1974.
39. Yager RR. Applications and extensions of OWA aggregations. Int J Man Mach Stud 1992;37:103–122.
40. Yager RR. A new methodology for ordinal multiple aspect decisions based on fuzzy sets, Dec Sci 1981;12:589–600.
41. Murphy PM, Aha DW. UCI Repository machine learning databases. Irvine, CA: University of California, Department of Information and Computer Science. 1994. Available at: http://www.ics.uci.edu/mlearn/MLRepository.html.
42. Data Extraction System. U.S. Census Bureau. Available at: http://www.census.gov/DES/www/welcome.html.
43. Torra V. On the re-identification of individuals described by means of non-common variables: a first approach. In: Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, Skopje, The Former Yugoslav Republic of Macedonia, March 2001. pp 14–16.
44. Domingo-Ferrer J, Torra V, Mateo-Sanz JM, Sebé F. Empirical disclosure risk assessment of the IPSO synthetic data generators. In: UNECE Work Sesison on Statistical Confidentiality, Geneve, Switzerland. 2005.