STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

COMMISSION OF THE EUROPEAN COMMUNITIES

CONFERENCE OF EUROPEAN STATISTICIANS

EUROSTAT

Joint ECE/Eurostat Work Session on Statistical Data Confidentiality
(Skopje, The former Yugoslav Republic of Macedonia, 14-16 March 2001)

Working Paper No. 17
English only

Topic II: Impact of new technological developments in software, communications and computing on SDC

ON THE RE-IDENTIFICATION OF INDIVIDUALS DESCRIBED BY MEANS OF NON-COMMON VARIABLES: A FIRST APPROACH

Contributed paper

Submitted by the Institut d'Investigació en Intelligència Artificial, Spain [1]

Abstract: In this paper we consider a first attempt to deal with re-identification of individuals when the variables in the two data files to be matched are not exactly the same but "similar".

Keywords: Statistical Disclosure Control, Re-identification procedures, Record linkage, Data mining

I. INTRODUCTION

1. The main task of National Statistical Offices (NSO) is to collect information from individuals and organizations and disseminate this information for researchers, media and the general public. The dissemination of the information is sometimes problematic due to disclosure risk. Disclosure risk is defined as the risk of re-identification of particular individuals. That is, some sensitive and confidential data that have been released are afterwards identified with a particular individual and, thus, confidentiality is lost. To avoid re-identification, data is distorted before its release. In this way, disclosure risk decreases. However, data has to maintain [Winkler, 1995] the so-called analytical validity, this is, data after being distorted have to reproduce the statistical analysis that can be produced with the original confidential data.

2. To evaluate the re-identification risk, re-identification methods are applied. Among these methods, we underline record linkage. Record linkage is used to link records in separate files that relate to the same individual or household. These methods [Newcomber et al.,1959], [Winkler, 1995] are based on the presence of a set of common variables in both files. The main difficulty that faces all these methods is that a matching procedure among pairs of records is not always enough to establish the link between the records. As [Winkler, 1995] points out, "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". To work with partial agreement between records, several methods have been developed in the literature.

3. Nowadays, due to the fact that the amount of available information is large, the disclosure risk increases. One of the threats to re-identification is the case in which the information available and the released information is not exactly the same (not the same variables) but somehow "similar". In this case,

[1] Prepared by Vicenç Torra (vtorra@iiia.csic.es). The author is also indebted to Josep Domingo for helpful comments and suggestions.

re-identification can be made on the basis of the similar available information. For example, a case of "similar" information is when there exists a correlation between variables. A file containing the income of individuals is released and available information corresponds to a file with the amount of money spent through credit card operations.

4. In this paper we present a first attempt to deal with the re-identification problem when the data files do not share the variables. We introduce re-identification procedures based on clustering algorithms to deal with this situation. We will present the basic assumptions we require for the re-identification, the formalization of the problem and some very preliminary results in this area.

II. CLUSTERING BASED RE-IDENTIFICATION

5. Re-identification for files with non-common variables is based on the existence of some basic information that is kept "constant" through files. We call this information "structural information". So, structural information (of data files) stands to any organization of the data that allows us to make explicit the relationship between individuals. This structural information is obtained from the data files through manipulation of the data in the file (e.g. using clustering techniques or any other data analysis or data mining technique). The comparison of the structural information implicit in both files is what allows the system to link two records that correspond to the same individual.

6. In the approach presented here, we use partitions to represent structural information. Partitions obtained from data by means of clustering techniques make implicit the relation between individuals according to the variables that describe them. Common partitions in both files correspond to the common structural information. We use partitions instead of other more sophisticated structures (like dendrograms) also obtainable from cluster methods because the former are less sensitive to changes in the data. Therefore, they can lead to better results. Results on consensus of classifications [Neumann, 1986] support this approach.

7. According to this, the following basic assumptions are made in this work:

Hypothesis 1. Both files share a large set of common individuals.
Hypothesis 2. Data in both files contain, implicitly, similar structural information.
Hypothesis 3. Structural information can be expressed by means of partitions.

8. To ease the re-identification process, instead of considering a single clustering technique, we apply several of them (different ones or the same ones but changing parameters) to both files. In this way, we obtain for each file and each technique a partition of the individuals. This initial process is formalized below considering that data files are named A and B, and as usual, files are defined by a set of records that assign values to variables. We assume that the file with known individuals is the file B.

9. Let $CP = \{CP_1, ..., CP_t\}$ be the set of clustering techniques considered. Then if $O_A=\{O^A_1, ..., O^A_{m(A)}\}$ and $O_B=\{O^B_1, ..., O^B_{m(B)}\}$ are the individuals in files A and B, then $CP_i(O^D_j)$ corresponds to the partition element that the i-th clustering technique assigns to the j-th object in file D.

10. To put both files into correspondence, we need to associate to each cluster in one data file to a cluster in the other one. However, as association has to be made cluster to cluster, a mapping for each clustering technique is needed. Therefore, we consider for each clustering technique CP_i, a function f_i that assigns a cluster in file B (one corresponding to CP_i) to each of the clusters in file A (one corresponding to CP_i). These functions have to be in a way that when applied to the clusters in file A, they return clusters as similar as possible to those in the second file. Based on this assumption, we have used the following similarity function:
$S(O^A= (x_1, x_2, ..., x_t), O^B= (y_1, y_2, ..., y_t) = \Sigma\, d_i(x_i, y_i)$,
where $d_i(x_i, y_i) = 1$ if $x_i= y_i$. and 0 otherwise.

11. To finish the formalization of the re-identification process we need the re-identification function. This is, a function that relates individuals of files A and B. We call this function n and takes the form: $n: \{1, ..., m(A)\} \rightarrow \{1, ..., m(B)\}$

12. Putting all this together, the re-identification problem can be formulated in the following way:

The re-identification problem: Find functions $f=(f_1, ..., f_t)$ and m such that the

$$\Sigma_{i=1, ,m(A)} S(f(CP_i(O^A_i)), O^B_{m(i)})$$

is maximized.

II.1 Example

13. In the following we show the feasibility of the approach, analyzing a small artificial problem. The example, that uses publicly available data, details all the steps described in the previous section.

14. To test the methodology with a well-known and public data file, we have considered the ionosphere database in the UCL repository [Murphy et al., 1994]. This example consists of a set of 351 examples (positive and negative examples) each defined in terms of 34 numerical variables.

15. To use this data for re-identification, two alternatives were possible: re-identification of the examples and re-identification of the variables. In the first alternative, the original file would be split so that all examples were present in both files, but only half of the variables would be present in each file. In the second alternative, the original file would be split so that all variables were present in both files but, instead, only half of the examples. We have followed the latter approach because it is not sure that half of the variables have enough information about the examples to allow for re-identification. Instead, two randomly chosen subsets of about 175 examples should give enough information about the structure of the variables. In fact, subsets of these examples are usually used in machine learning [Sigillito et al., 1989]. They assume that subsets have still enough information on the variables. In other words, this approach was used because we assumed more redundancy in the examples than in the variables.

16. To apply the method described above, we have considered an initial normalization step. It consisted on the normalization of the domain of the variables and it was applied before the file was partitioned (usual normalization in the [0,1] interval was applied: x' = (x - min) / (max - min)). Then, the set of examples was randomly divided into two sets. One set resulted with 170 and the other with 181 examples.

Table 1. Partitions obtained for file A and two re-identification functions: m1 and m2 are the re-identification functions; the last six columns to partitions. In the last six columns, aa and cc correspond to the classification criteria and stand for Arithmetic average and Centroid clustering; m, d and t refer to similarity functions based, respectively, on Manhattan distance, Differences and Taxonomic distance.

Π(A,CP)	m1	m2	aa,m	aa,d	aa,t	cc,m	cc,d	cc,t
O^A_1	1	1	$c_{1,A,4}$	$c_{2,A,4}$	$c_{3,A,1}$	$c_{4,A,1}$	$c_{5,A,3}$	$c_{6,A,3}$
O^A_2	2	2	$c_{1,A,1}$	$c_{2,A,2}$	$c_{3,A,4}$	$c_{4,A,4}$	$c_{5,A,3}$	$c_{6,A,5}$
O^A_4	3	3	$c_{1,A,2}$	$c_{2,A,2}$	$c_{3,A,2}$	$c_{4,A,2}$	$c_{5,A,6}$	$c_{6,A,2}$
O^A_5	4	4	$c_{1,A,4}$	$c_{2,A,2}$	$c_{3,A,1}$	$c_{4,A,3}$	$c_{5,A,6}$	$c_{6,A,1}$
O^A_6,O^A_8	6	6	$c_{1,A,2}$	$c_{2,A,2}$	$c_{3,A,2}$	$c_{4,A,2}$	$c_{5,A,4}$	$c_{6,A,2}$
O^A_7,O^A_9	5	5	$c_{1,A,4}$	$c_{2,A,2}$	$c_{3,A,1}$	$c_{4,A,3}$	$c_{5,A,4}$	$c_{6,A,1}$
O^A_{10}	10	10	$c_{1,A,2}$	$c_{2,A,1}$	$c_{3,A,2}$	$c_{4,A,2}$	$c_{5,A,3}$	$c_{6,A,2}$
O^A_3,O^A_{11}	7	7	$c_{1,A,4}$	$c_{2,A,2}$	$c_{3,A,1}$	$c_{4,A,3}$	$c_{5,A,3}$	$c_{6,A,1}$
O^A_{14}	10	10	$c_{1,A,2}$	$c_{2,A,1}$	$c_{3,A,2}$	$c_{4,A,2}$	$c_{5,A,5}$	$c_{6,A,2}$
O^A_{15}	14	14	$c_{1,A,4}$	$c_{2,A,1}$	$c_{3,A,1}$	$c_{4,A,3}$	$c_{5,A,5}$	$c_{6,A,1}$
$O^A_{12}O^A_{16}$	11	11	$c_{1,A,2}$	$c_{2,A,1}$	$c_{3,A,2}$	$c_{4,A,2}$	$c_{5,A,4}$	$c_{6,A,2}$
O^A_{18}	16	13	$c_{1,A,3}$	$c_{2,A,1}$	$c_{3,A,2}$	$c_{4,A,2}$	$c_{5,A,1}$	$c_{6,A,2}$
O^A_{19}	17	21	$c_{1,A,4}$	$c_{2,A,1}$	$c_{3,A,1}$	$c_{4,A,3}$	$c_{5,A,1}$	$c_{6,A,1}$
O^A_{20}	15	13	$c_{1,A,3}$	$c_{2,A,1}$	$c_{3,A,2}$	$c_{4,A,2}$	$c_{5,A,2}$	$c_{6,A,2}$
O^A_{21}	21	21	$c_{1,A,4}$	$c_{2,A,1}$	$c_{3,A,1}$	$c_{4,A,3}$	$c_{5,A,3}$	$c_{6,A,1}$
$O^A_{22}O^A_{24}$	18	18	$c_{1,A,3}$	$c_{2,A,1}$	$c_{3,A,2}$	$c_{4,A,2}$	$c_{5,A,4}$	$c_{6,A,2}$
O^A_{25}	20	21	$c_{1,A,4}$	$c_{2,A,1}$	$c_{3,A,1}$	$c_{4,A,3}$	$c_{5,A,2}$	$c_{6,A,1}$
$O^A_{13}O^A_{17}$, $O^A_{23}O^A_{27}$	19	17	$c_{1,A,4}$	$c_{2,A,1}$	$c_{3,A,1}$	$c_{4,A,3}$	$c_{5,A,4}$	$c_{6,A,1}$
O^A_{28}	22	22	$c_{1,A,3}$	$c_{2,A,3}$	$c_{3,A,2}$	$c_{4,A,2}$	$c_{5,A,6}$	$c_{6,A,2}$

$O^A_{29}O^A_{31}$	19	19	$c_{1,A,4}$	$c_{2,A,3}$	$c_{3,A,1}$	$c_{4,A,3}$	$c_{5,A,4}$	$c_{6,A,1}$
$O^A_{26}O^A_{30}$, O^A_{32}	22	22	$c_{1,A,3}$	$c_{2,A,3}$	$c_{3,A,2}$	$c_{4,A,2}$	$c_{5,A,4}$	$c_{6,A,2}$
O^A_{33}	21	21	$c_{1,A,4}$	$c_{2,A,3}$	$c_{3,A,1}$	$c_{4,A,3}$	$c_{5,A,6}$	$c_{6,A,1}$
O^A_{34}	22	22	$c_{1,A,3}$	$c_{2,A,3}$	$c_{3,A,2}$	$c_{4,A,2}$	$c_{5,A,7}$	$c_{6,A,2}$
O^A_{35}	23	23	$c_{1,A,4}$	$c_{2,A,3}$	$c_{3,A,3}$	$c_{4,A,1}$	$c_{5,A,4}$	$c_{6,A,4}$

17. The next step was to obtain the partitions. To do so, six classification techniques were applied to both files. Each technique led to a dendrogram, and for each dendrogram a partition was obtained. Dendrograms were obtained using SAHN [Everitt, 1977] methods (i.e., sequential, agglomerative, hierarchic, non-overlapping methods). Different selection of similarity functions (functions to compute similarities between objects/classes) and of classification criteria (how to compute, from already known similarities, the similarity between a new class and the previous existing ones) were applied. Three similarity functions were used (based on Manhattan distance, Differences and Taxonomic distance) combined with two classification criteria (Arithmetic average, centroid clustering). The definition of these functions and their properties are detailed in [Everitt, 1977]. The partitions obtained by the six classification methods are given in Tables 1 and 2.

18. With all this information, we have solved the maximization problem formalized above. Two solutions are given. One restricting f_i to be a one-to-one function and another one allowing f_5 to be such that $f_5(x)=f_5(y)$ when $x \neq y$. The functions f_i of the solution are given in Table 3 while the group level re-identification functions are given in Table 1. In this latter table, m1 correspond to the function with f_5 being a one-to-one function while m2 corresponds to the other case.

19. Once the method has been applied, we have compared its performance with respect to correct re-identifications. To do so, we have computed the number of objects that belong to the class pointed out by the re-identification function (m1 and m2, respectively). This has been normalized by the number of elements so that the function has a maximum value of 1. This is:

$$| \{O^A_i \mid O^A_i \in \pi^B_{m(i)}\} | / | O^A |$$

where m is the re-identification function, and π is the set of all elements that are equivalent in relation to the partitions in B (e.g., $\pi^6=\{O^B_8, O^B_{10}\}$). It can be observed that in both cases, 20 out of 35 objects have been correctly re-identified. Thus, 57% of the objects have been correctly re-identified. In both case the similarity function S between the two partitions is 137. Correctly re-identified objects are not the same for both cases.

Table 2. Partitions obtained for file B using the same methods that in table 1.

		aa,m	aa,d	aa,t	cc,m	cc, d	cc, t
1	O^B_1	$c_{1,B,1}$	$c_{2,B,4}$	$c_{3,B,1}$	$c_{4,B,3}$	$c_{5,B,5}$	$c_{6,B,3}$
2	O^B_2	$c_{1,B,4}$	$c_{2,B,2}$	$c_{3,B,4}$	$c_{4,B,4}$	$c_{5,B,5}$	$c_{6,B,5}$
3	O^B_4, O^B_6	$c_{1,B,2}$	$c_{2,B,2}$	$c_{3,B,3}$	$c_{4,B,1}$	$c_{5,B,2}$	$c_{6,B,2}$
4	O^B_5, O^B_7	$c_{1,B,1}$	$c_{2,B,2}$	$c_{3,B,1}$	$c_{4,B,2}$	$c_{5,B,2}$	$c_{6,B,1}$
5	O^B_9	$c_{1,B,1}$	$c_{2,B,2}$	$c_{3,B,1}$	$c_{4,B,2}$	$c_{5,B,1}$	$c_{6,B,1}$
6	O^B_8, O^B_{10}	$c_{1,B,2}$	$c_{2,B,2}$	$c_{3,B,3}$	$c_{4,B,1}$	$c_{5,B,1}$	$c_{6,B,2}$
7	O^B_3, O^B_{11}	$c_{1,B,1}$	$c_{2,B,2}$	$c_{3,B,1}$	$c_{4,B,2}$	$c_{5,B,5}$	$c_{6,B,1}$
8	O^B_{12}	$c_{1,B,2}$	$c_{2,B,2}$	$c_{3,B,3}$	$c_{4,B,1}$	$c_{5,B,7}$	$c_{6,B,2}$
9	O^B_{13}	$c_{1,B,1}$	$c_{2,B,2}$	$c_{3,B,1}$	$c_{4,B,2}$	$c_{5,B,3}$	$c_{6,B,1}$
10	O^B_{14}	$c_{1,B,2}$	$c_{2,B,1}$	$c_{3,B,3}$	$c_{4,B,1}$	$c_{5,B,3}$	$c_{6,B,2}$
11	O^B_{16}	$c_{1,B,2}$	$c_{2,B,1}$	$c_{3,B,3}$	$c_{4,B,1}$	$c_{5,B,1}$	$c_{6,B,2}$
12	O^B_{20}	$c_{1,B,3}$	$c_{2,B,1}$	$c_{3,B,3}$	$c_{4,B,1}$	$c_{5,B,3}$	$c_{6,B,2}$
13	O^B_{18}, O^B_{22}	$c_{1,B,3}$	$c_{2,B,1}$	$c_{3,B,3}$	$c_{4,B,1}$	$c_{5,B,2}$	$c_{6,B,2}$
14	O^B_{15}, O^B_{19}, O^B_{23}	$c_{1,B,1}$	$c_{2,B,1}$	$c_{3,B,1}$	$c_{4,B,2}$	$c_{5,B,3}$	$c_{6,B,1}$
15	O^B_{24}	$c_{1,B,3}$	$c_{2,B,1}$	$c_{3,B,3}$	$c_{4,B,1}$	$c_{5,B,4}$	$c_{6,B,2}$
16	O^B_{26}	$c_{1,B,3}$	$c_{2,B,1}$	$c_{3,B,3}$	$c_{4,B,1}$	$c_{5,B,6}$	$c_{6,B,2}$
17	O^B_{27}	$c_{1,B,1}$	$c_{2,B,1}$	$c_{3,B,1}$	$c_{4,B,2}$	$c_{5,B,6}$	$c_{6,B,1}$
18	O^B_{28}	$c_{1,B,3}$	$c_{2,B,1}$	$c_{3,B,3}$	$c_{4,B,1}$	$c_{5,B,1}$	$c_{6,B,2}$
19	O^B_{17}, O^B_{29}	$c_{1,B,1}$	$c_{2,B,1}$	$c_{3,B,1}$	$c_{4,B,2}$	$c_{5,B,1}$	$c_{6,B,1}$
20	O^B_{25}, O^B_{31}	$c_{1,B,1}$	$c_{2,B,1}$	$c_{3,B,1}$	$c_{4,B,2}$	$c_{5,B,4}$	$c_{6,B,1}$
21	O^B_{21}, O^B_{33}	$c_{1,B,1}$	$c_{2,B,1}$	$c_{3,B,1}$	$c_{4,B,2}$	$c_{5,B,2}$	$c_{6,B,1}$

22	$O^B_{30}, O^B_{32}, O^B_{34}$	$c_{1,B,3}$	$c_{2,B,3}$	$c_{3,B,3}$	$c_{4,B,1}$	$c_{5,B,2}$	$c_{6,B,2}$
23	O^B_{35}	$c_{1,B,1}$	$c_{2,B,3}$	$c_{3,B,2}$	$c_{4,B,3}$	$c_{5,B,2}$	$c_{6,B,4}$

Table 3. Functions f_i to maximize the similarity between partitions in Tables 4 and 5.

$f_1(c_1\text{-A})$	$f_2(c_2\text{-A})$	$f_3(c_3\text{-A})$	$f_4(c_4\text{-A})$	$f_5(c_5\text{-A})$	$f'_5(c_5\text{A})$	$f_6(c_6\text{-A})$
c1-B-4	c2-B-1	c3-B-1	c4-B-3	c5-B-6	c5-B-2	c6-B-1
c1-B-2	c2-B-2	c3-B-3	c4-B-1	c5-B-4	c5-B-2	c6-B-2
c1-B-3	c2-B-3	c3-B-2	c4-B-2	c5-B-5	c5-B-5	c6-B-3
c1-B-1	c2-B-4	c3-B-4	c4-B-4	c5-B-1	c5-B-1	c6-B-4
				c5-B-3	c5-B-3	c6-B-5
				c5-B-2	c5-B-2	
				c5-B-7	c5-B-2	

Table 4. Probabilities of having r correct links, and of having more or equal than r links for 33 individuals.

0	0.36787942	1.0
1	0.36787942	0.63212055
2	0.18393971	0.26424113
3	0.06131324	0.0803014
4	0.01532831	0.018988157
5	0.003065662	0.0036598467
6	5.109437E-4	5.941848E-4
7	7.299195E-5	8.3241146E-5
8	9.123994E-6	1.0249197E-5
9	1.0137771E-6	1.1252026E-6
10	1.01377715E-7	1.1142548E-7
11	9.216156E-9	1.0047766E-8
12	7.680129E-10	8.316107E-10
13	5.907792E-11	6.359777E-11
14	4.2198515E-12	4.5198524E-12
15	2.8132344E-13	3.0000107E-13
16	1.7582715E-14	1.8677634E-14
17	1.0342773E-15	1.0949201E-15
18	5.745985E-17	6.064281E-17
19	3.0242027E-18	3.1829554E-18
20	1.5121014E-19	1.5875276E-19
21	7.200482E-21	7.5426254E-21
22	3.2729465E-22	3.4214245E-22
23	1.4230203E-23	1.4847793E-23
24	5.929247E-25	6.1758905E-25
25	2.3717165E-26	2.4664351E-26
26	9.121372E-28	9.471847E-28
27	3.3801082E-29	3.5047438E-29
28	1.202626E-30	1.246356E-30
29	4.241236E-32	4.3729944E-32
30	1.2566625E-33	1.317584E-33
31	6.080625E-35	6.092142E-35
32	0	1.1516336E-37
33	1.1516335E-37	1.1516336E-37

II.2 Evaluation of the results

20. To evaluate the results obtained by the clustering method, we have computed the probability that taking an assignment from file A to file B at random, there are more than x correct re-identified elements. To compute this probability, note that when m(A) = m(B) = N, and $n(i) \neq n(j)$ for $i \neq j$, then n can be seen as a permutation. Then, to have r correct re-identifications, corresponds to a permutation such that r elements satisfy n(i) = i.

21. These permutations can be built in the following way (we use below k=N-r): We take k elements from the correct permutation (i.e., p(i)=i for all i) and we permutate the k elements in such a way that there is no one that keeps its original position. To compute the number of these permutations consider that the number of possible k elements to be taken is n! / (k! (N-k)!); and that *permutations without fixed point* (see e.g. [Reinhardt et al., 1997]) correspond to the number of permutations of k elements in such a way that there is no one that keeps its original position. The number of permutations of k elements without fixed point is [Reinhardt et al., 1997]:

$$pwfp(k) = k! \sum_{v=0}^{k} (-1)^k / v!$$

Therefore, the number of permutations such that there are exactly r elements in the correct position is:

$$(N! \sum_{v=0}^{k} (-1)^k / v!)/ (N-k)!$$

Now, as the number of total permutations (the number of all possible re-identifications) is N!, the probability of finding at random a permutation with exactly r elements in the correct position is:

$$(\sum_{v=0}^{k} (-1)^k / v!)/ (N-k)!$$

22. We give in Table 4, the corresponding probabilities. In fact, we give the probabilities for N=33 because while the original file has 34 variables, variable 2 is always zero and thus its re-identification is rather simple. Also, the last variable of the file is generated from the other ones. Thus, the proper probability is the one for k=19 re-identifications over 33 variables. Note that this probability is about $3 \cdot 10^{-18}$. This value show that although re-identification for files with non-common variables is far from optimal it is possible in some extend.

III. CONCLUSIONS AND FUTURE WORK

23. The results given in this work are a first attempt to deal with the problem of re-identification of individuals when non-common variables are shared by two information sources. The results obtained show that it is feasible to deal with this problem, although more extensive research is needed in this direction. In particular, we can point out the following topics:

- Exhaustive testing. We have only tested a small example with some particular methods. Validity should be assessed with exhaustive testing with large data files and several clustering techniques.
- Analyze which of the clustering techniques are more suitable for re-identification. This need is already suggested by the example described. For example, note that for one of the clustering methods (i.e., centroid clustering with differences based similarity) there is no correlation between partitions in files A and B.
- Identify appropriate algorithms for group level identification to increase the effectiveness of the system: to obtain and assure the optimal for a particular problem and to compute the optimal with as less partitions as possible.
- Extension to non-numerical data. The approach introduced here has been applied to quantitative data, however the same approach can be applied to data files with other types of data (as ordinal – qualitative [Godo et al., 2000]) when clustering techniques exists that deals with this data.

Also, techniques other than clustering can be used to represent structural information (e.g., aggregation operators as in [Torra, 2000]).

References

[1] J. Domingo, V. Torra, On the Connections Between Statistical Disclosure Control for Microdata and Some Artificial Intelligence Tools, (submitted).
[2] Everitt, B. (1977), Cluster analysis, Heinemann Educational Books Ltd.
[3] Godo, L., Torra, V., On aggregation operators for ordinal qualitative information, *IEEE T. Fuzzy Systems,(in press)*.
[4] 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.
[5] Neumann, D.A., Norton, V.T., (Jr), (1986), Clustering and isolation in the consensus problem for partitions, *Journal of classification*, 3 281-297.

[6] H. B. Newcombe, J. M. Kennedy, S. J. Axford, A. P. James, (1959), Automatic Linkage of Vital Records, Science, 130, 954--959.
[7] Reinhard, F., Soeder, H., (1997), Atlas des mathématiques, Librairie Générale Française.
[8] Sigillito, V. G., Wing, S. P., Hutton, L. V., \& Baker, K. B. (1989), Classification of radar returns from the ionosphere using neural networks. Johns Hopkins APL Technical Digest, 10, 262-266.
[9] Torra, V., (2000), Towards the Re-identification of Individuals in Data Files with Non-common Variables, European Conference on Artificial Intelligence, IOS press, Berlin, Germany.
[10] Torra, V., (2000), Re-identifying individuals using OWA operators, Proc. of the Iizuka conference, IIZUKA, Japan (CD-ROM).
[11] W.E. Winkler, (1995), Matching and Record Linkage, in B. G. Cox (ed.), Business Survey Methods, J. Wiley, 355-384.