Towards the use of OWA operators for record linkage

EUSFLAT - LFA 2005

Jordi Nin
IIIA-CSIC
Campus UAB s/n
08193 Bellaterra (Catalonia, Spain)
e-mail: jnin@iiia.csic.es

Vicenç Torra
IIIA-CSIC
Campus UAB s/n
08193 Bellaterra (Catalonia, Spain)
e-mail: vtorra@iiia.csic.es

Abstract

Record linkage is used to establish links between those records that while belonging to two different files correspond to the same individual. Classical approaches assume that the two files contain some common variables, that are the ones used to link the records.

Recently, we introduced a new approach to link records among files when such common variables are not available. In this approach, re-identification is based on the so-called structural information. In this paper we study the use of OWA operators for extracting such structural information and, thus, allowing re-identification.

Keywords: Record linkage, OWA operators, data mining, data cleaning, privacy preserving data mining.

1 Introduction

In recent years, due to the ease in information gathering, the amount of information stored in mass storage system for any individual has increased dramatically [6]. At the same time, the ubiquitous presence of computers causes that this information is mainly distributed and represented in an heterogeneous way.

Due to this, the importance of tools for data cleaning [11] and integration has increased. Re-identification algorithms are one of such tools. They are used to identify the structures that are shared by several files or databases.

Record linkage algorithms (see reviews in [7], [11]) are one of the most important re-identification tools. Their goal is establish which records give information on the same individual. For example, we can consider two files, one corresponding to providers of a company and the other corresponding to customers of the same company. Record linkage can be used to find those individuals that are at the same time providers and customers. Record linkage algorithms are used for several different purposes. Its main use is data integration. In this case, they are used in conjunction with data consolidation methods. Another use is for risk assessment in privacy preserving data mining (PPDM) or statistical disclosure control (SDC). In this setting, they permit to evaluate whether a protection mechanism provides enough protection to providers of sensitive information (no disclosure can be guaranteed).

Classical record linkage methods focus on the linkage of records from two files when such files share a set of variables. In this case, the difficulties for a good performance of record linkage algorithms are due to the fact that files contain errors (*e.g.*, the salary of an individual is not the same in both files). These errors [1] may be accidental (*e.g.*, due to incorrect manipulation of data) or intentional (*e.g.* to protect sensitive information as in PPDM).

In recent works [9, 10], we showed that re-identification is also possible in other situations in which files do not share a set of variables. We proved [2] that re-identification is a threat to privacy preserving data mining as records can be re-identified even when no common variables are shared by two (or more) files.

In general, re-identification is possible when the following set of assumptions hold:

Assumption 1 : Both files share a large set of common individuals.

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

When different formalisms are considered for representing the structural information, we can develop different methods for re-identification.

In this work we consider the use of aggregation operators as the basic brick for re-identification. This is based on the idea of using aggregation operators for building summaries [13] from data. In this paper, we assume that the structural information is expressed in terms of some numerical representatives, and that numerical representatives are extracted from the records using aggregation operators. This approach is formalized adding the following two assumptions to the previous ones:

Assumptions 3 : Structural information can be expressed by means of numerical representatives for each individual.

Assumptions 4 : Aggregation operators are used as the summarization mechanisms for each individual.

Note that the first assumption implies that re-identification is possible (there are records to link). The second assumption is to say that there are some similarities between the different individuals that are kept more or less constant and do not depend on the files. We call these similarities *structural information*. The third assumption and the fourth one are the ones that justify the use of aggregation procedures for re-identification.

In this work we develop the approach of using aggregation operators for record linkage when files do not share variables. In particular, we consider the use of OWA operators for this task. A previous work [8, 10] showed that this approach was appropriate for re-identification of variables. We extend here our previous work to the re-identification of records. It has to be said that the approach is similar in both cases, eventhough in the case of re-identification of variables the results might be better as the data related to variables are more redundant than the one on records. The non exportability of the results was pointed out in [5, 4]). Thus, the experiments reported here on records are more conclusive than the previous ones reported in [8, 10].

The structure of the paper is as follows. In Section 2 we describe some elements that are needed latter on. Then, in Section 3 we introduce our approach to record linkage. Section 4 describes some of the experiments performed. Then, the paper finishes with some conclusions and description of future work.

2 Preliminaries

In this section we review a few definitions that are needed latter on. We start with the definion of the OWA operator in terms of a fuzzy quantifier [3].

Definition 1 *A function $Q : [0,1] \to [0,1]$ is a* regular monotonically non-decreasing fuzzy quantifier *(non-decreasing fuzzy quantifiers for short) if it satisfies: (i) $Q(0) = 0$; (ii) $Q(1) = 1$; (iii) $x > y$ implies $Q(x) \geq Q(y)$.*

Definition 2 *[12] Let Q be a non-decreasing fuzzy quantifier, then a mapping $OWA_Q : \mathbb{R}^N \to \mathbb{R}$ is an* Ordered Weighting Averaging (OWA) operator *of dimension N if*

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

where σ is defined as a permutation of $\{1, \ldots, N\}$ such that $a_{\sigma(i)} \geq a_{\sigma(i+1)}$.

2.1 Re-identification methods

As said in the introduction, the goal of classical re-identification methods is to link records in two files that correspond to the same individual and that are described using the same variables. Two main approaches have been defined for this purpose (see [7] for details and references):

Distance-based Record Linkage: Records of two files A and B are compared, and each

record in A is linked to the nearest record in B.

Probabistic Record Linkage: Conditional probabilities of coincidence (and non-coincidence) of values among records given correct matching are obtained. From these conditional probabilities an index is computed for each pair of records (a,b) with a in A and b in B. This index is used to classify pairs as linked (a and b correspond to the same individual).

2.2 Record linkage evaluation

To evaluate the performance of a record linkage, we should compare its effectiveness with alternative approaches. Nevertheless, being a new approach, we compare our method with the probability of re-identification using a random strategy. Such probability is defined in the next proposition. Tables 1 and 2 display some of the values for this probability when the number of records are 100 and 30, respectively.

Proposition 1 *[2, 10] 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*

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

3 Our approach to record linkage

As pointed out in the introduction, our objective is the development of a method for record linkage for files not sharing variables, and using aggregation operators as the basic tool for extracting the structural information.

The use of OWA operators based on quantifiers was motivated on the fact that:

- They are aggregation operators that can be used for aggregating an arbitrary number of values. This is useful here as different files can contain different number of variables.
- They are symmetric. This is suitable as there are not *a priori* variables with a known larger importance.

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
4	0.01532831	0.01898816
5	0.00306566	0.00365985
6	0.00051094	0.00059418
7	0.00007299	0.00008324
8	0.00000912	0.00001025
9	0.00000101	0.00000113
10	1.01378E-7	1.11425E-7
15	1.5121E-19	1.5875E-19
20	2.3717E-26	2.4664E-26
30	1.3869E-33	1.4331E-33
40	4.5088E-49	4.6214E-49
50	1.2096E-65	1.2338E-65
75	1.483E-110	1.503E-110
100	1.071E-158	1.071E-158

Table 1: Probabilities of having r correct links, and of having more or equal than r links for 100 records.

- They are parametric. This is appropriate here as several different representatives can be built from the same data using different parameters.

Considering the OWA operator and each parameterization as a method to build a representative, we have that several numerical representatives can be built from each record. When the same approach *i.e.*, the same parameterization, is applied to two files of records, we obtain two files of representatives. This is illustrated in Figure 1. Let A and B be the original files, then the new files A' and B' contain the same number of records than the original files A and B and contain the same number of variables. Note that the number of variables in both files A' and B' corresponds to the number of parameterizations considered.

Once the two new files A' and B' are built, it is possible to apply re-identification algorithms to link them. In this case, as both files have been constructed using the same parameters, we consider them as defined in terms of the same variables (in our case, one variable is one parameteri-

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
4	0.01532831	0.01898816
5	0.00306555	0.00365985
6	0.00051094	0.00059418
7	0.00007299	0.00008324
8	0.00000912	0.00001025
9	0.00000101	0.00000113
10	1.01378E-7	1.11425E-7
15	2.8132E-13	3.0000E-13
20	1.5121E-19	1.5875E-19
25	2.3639E-26	2.4601E-26
30	3.7670E-33	3.7670E-33

Table 2: Probabilities of having r correct links, and of having more or equal than r links for 30 records.

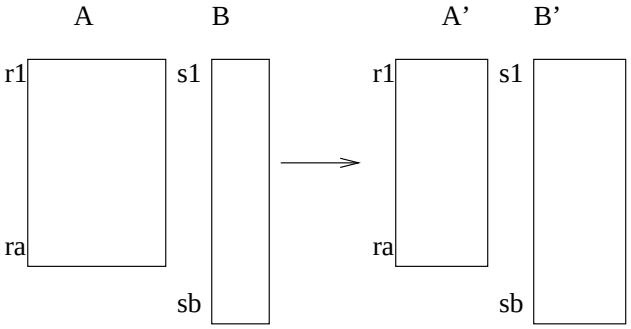

Figure 1: Graphical representation of our approach to re-identification using OWA operators

zation). Therefore, we can apply standard record linkage approaches (distance-based record linkage and probabilistic record linkage).

As *a priori*, it was not known what is the best parameterization in terms of the number of reidentifications, we have considered different sets of parameters.

4 Experiments

To analyze the feasibility of our approach we have studied seven different problems. These problems have been generated from data publicly available from the UCL repository. In particular, we have considered the following files: abalone, iris, ionosphere, dermatology, housing, water-treatment, wdbc. In addition, we also used publicly available data from the U.S. Census Bureau: "1995 March Questionnaire Supplement-Person Data Files".

To test the reidentification approach, we have divided the datafiles from these repositories into two parts. Namely, records are divided into two parts, separating the variables with higher correlation. In this way, we obtain two files with the same records but where such records are described in terms of different variables. The reason for separating the data according to the correlation coefficient is that in this way, it might be considered that the two files contain similar information. Some variables have been discarded if they are not correlated with the others (when all correlation coefficients were less than 0.7 the variable was removed). Additionally, records with missing values were not considered for re-identification.

Test files of different size (different number of records) have been considered to evaluate the method performance with respect to size. In particular, we have used files of 30 and 100 records.

Before the application of our approach we have included a pre-processing step that consisted on the normalization/standardization of the data. The two following alternatives have been considered:

Ranging: Translation of data values from the $[\max, \min]$ interval into [0,1] using $x' = (x - \min(v))/(\max(v) - \min(v))$ (where x is the previous value, and $\max(v)$ and $\min(v)$ are the maximum and minimum values for the corresponding variable v).

Standardization: Mean equals zero and both the standard deviation and variance equals one: $x = (x - \bar{v})/\sigma$ (where $\bar{v}$ and $\sigma(v)$ are, respectively, the mean and the standard deviation of the corresponding variable v).

In our experiments, we have tested three different fuzzy quantifiers. The quantifiers are defined below and their graphical representation is given in Figures 2, 3 and 4:

$Q_1^\alpha(x) = x^\alpha$ for $\alpha = 1/5, 2/5, \cdots, \ldots, 10/5$

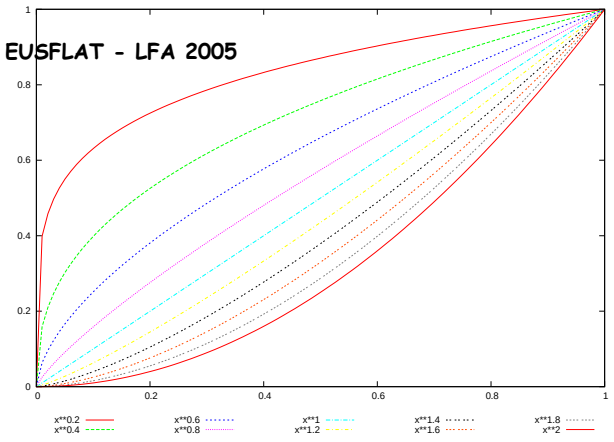

Figure 2: Graphical representation of Q_1^α

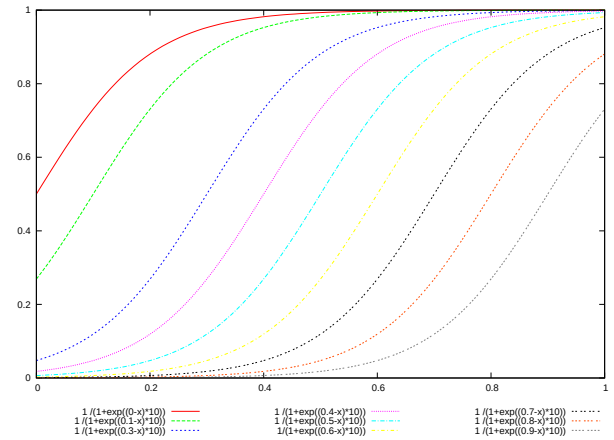

Figure 3: Graphical representation of Q_2^α

$$Q_2^\alpha(x) = 1/(1+e^{(\alpha-x)*10}) \text{ for } \alpha = \{0, 0.1, \ldots 0.9\}$$

$$Q_3^\alpha(x) = \begin{cases} 0 & \text{if } x \leq \alpha \\ 1 & \text{if } x > \alpha \end{cases} \quad \text{for } \alpha = \{0, 0.1, \ldots 0.9\}$$

The results we have obtained using our approach with these quantifiers are summarized in the following tables:

To evaluate in an appropriate way the performance of our approach and the results displayed in Tables 3, 4 and 5 we compare the correct links (hits) achieved and the probability of obtaining such results using a random approach. This latter probability was defined in Proposition 1.

If we compare the results obtained in the experiments made with the tables of random record linkage probabilities, we observe that the obtained results are good. I.e., with the ionosphere problem we have obtained 27 hits when we considered files with 100 records and 10 hits when files had 30 records. The probability to have these results at random is 3.37849E-29 in case of 100 records and 1.0138E-7 in case of 30.

We have also obtained good results in the wdbc problem, where we have obtained 26 hits from a file of 100 records (Probability at random equal to 9.1219E-28) and 16 hits from 30 (P = 1.75827E-14 at random), the water-treatment problem with 17 hits from 100 records (P = 1.0343E-15 at random) and 11 hits from 30 (P = 9.21616E-9 at random)

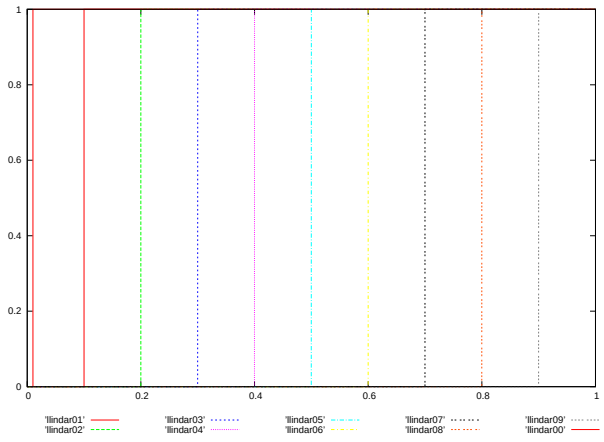

Figure 4: Graphical representation of Q_3^α

	R/P	R/D	S/P	S/D	R/P	R/D	S/P	S/D
abalone	3	6	13	15	2	6	8	8
census	1	7	9	8	6	5	5	4
iris	6	4	2	4	3	2	1	2
ionosphere	12	10	4	7	5	6	1	5
dermatology	3	3	0	0	0	4	1	1
housing	1	2	6	7	4	2	2	4
water-tr.	0	3	8	7	0	11	2	5
wdbc	5	6	10	10	5	4	7	7

Table 3: Results for 100 records and 30 records, quantifier Q_1^α. R/· stands for Ranging, S/· stands for Standardization, ·/D for Distance based record linkage and ·/P for probabilistic record linkage.

	R/P	R/D	S/P	S/D	R/P	R/D	S/P	S/D
abalone	7	10	13	10	1	2	1	10
census	9	10	8	9	4	5	4	5
iris	1	3	2	4	1	4	4	3
ionosphere	21	21	7	5	10	9	4	5
dermatology	0	0	1	1	1	0	1	3
housing	2	5	5	3	2	4	2	5
water-tr.	8	11	17	15	5	5	7	10
wdbc	19	10	14	17	4	3	11	11

Table 4: Results for 100 records and 30 records, quantifier Q_2^α. R/P $\cdots$ S/D is as in Table 3.

	R/P	R/D	S/P	S/D	R/P	R/D	S/P	S/D
abalone	2	8	2	7	3	7	2	11
census	4	4	4	11	3	3	4	5
iris	2	2	1	3	1	1	1	0
ionosphere	3	27	1	10	2	6	1	4
dermatology	1	1	1	0	8	7	1	5
housing	1	6	1	4	2	5	1	4
water-tr.	2	8	5	10	3	4	4	6
wdbc	9	4	19	26	6	7	16	15

Table 5: Results for 100 records and 30 records, quantifier Q_3^α. R/P $\cdots$ S/D is as in Table 3.

and the abalone problem with 15 hits from 100 records (P = 2.81323E-13 at random) and 11 hits from 30 (P = 9.21616E-7 at random).

5 Conclusions and future work

In this paper, we have studied an alternative method for record linkage based on structural information, and focused in the particular case in which information is numerical. We have proved that owa operators are a suitable tools for such re-identification as they have lead to good results in eight different problems.

In order to complete this work we have to make more experiments, use alternative quantifiers and consider the use of a combination of quantifiers from different families. Also, we need to refine our algorithm as the current method does not permit to extract easily the correct links from all the set of suggested links.

Acknowledgments

This work was partly funded by the Spanish Ministry of Education and Science under project SEG2004-04352-C04-02 "PROPRIETAS".

References

[1] Domingo-Ferrer, J., Torra, V., (2003), On the Connections Between Statistical Disclosure Control for Microdata and Some Artificial Intelligence Tools, Inf. Sci., 151 153-170.

[2] Domingo-Ferrer, J., Torra, V., (2003), Disclosure Risk Assessment in Statistical Microdata Protection via advanced record linkage, Statistics and Computing, 13 343-354.

[3] Klir, G., Yuan, B., (1995), Fuzzy Sets and Fuzzy Logic: Theory and Applications, Prentice-Hall, U.K.

[4] Malin, B., (2005), Betrayed By My Shadow: Learning Data Identity via Trail Matching, Journal of Privacy Technology (http://www.jopt.org).

[5] Malin, B., Sweeney, L., Newton, E., (2003), 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.

[6] Sweeney, L., (2001), Information explosion, in P. Doyle, J. I. Lane, J. M. Theeuwes and L. M. Zayatz (Eds.), Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies, Elsevier, 43–74.

[7] Torra, V., Domingo-Ferrer, J., (2003), Record linkage methods for multidatabase data mining, in V. Torra (Ed.), Information Fusion in Data Mining, Springer, 101-132.

[8] Torra, V., (2000), Re-identifying Individuals using OWA operators, Proc. of the 6th Int. Conference on Soft Computing (Iizuka 2000), (CD Rom), Iizuka, Fukuoka, Japan.

[9] Torra, V., (2000), Towards the re-identification of individuals in data files with Non-common variables, Proc. of the 14th European Conference on Artificial Intelligence (ECAI2000) (IOS Press, ISBN 1 58603 013 2), 326-330, Berlin, Germany.

[10] Torra, V., (2004), OWA operators in data modeling and re-identification, IEEE Trans. on Fuzzy Systems, 12:5 652-660.

[11] Winkler, W., (2003), Data Cleaning Methods, Proc. SIGKDD 2003, Washington.

[12] Yager, R. R., (1993), Families of OWA operators, Fuzzy Sets and Systems, 59 125-148.

[13] Yager, R. R., (2004), Data Mining Using Granular Linguistic Summaries, in V. Torra, Information Fusion in Data Mining, Springer, 211-229.