Statistical Perspective on Blocking Methods When Linking Large Data-sets

Nicoletta Cibella and Tiziana Tuoto

Abstract The combined use of data from different sources is largely widespread. Record linkage is a complex process aiming at recognizing the same real world entity, differently represented in data sources. Many problems arise when dealing with large data-sets, connected with both computational and statistical aspects. The well-know blocking methods can reduce the number of record comparisons to a suitable number. In this context, the research and the debate are very animated among the information technology scientists. On the contrary, the statistical implications of different blocking methods are often neglected. This work is focused on highlighting the advantages and disadvantages of the main blocking methods in carrying out successfully a probabilistic record linkage process on large data-sets, stressing the statistical point of view.

1 Introduction

The main purpose of record linkage techniques is to accurately recognize the same real world entity which can be differently stored in sources of various type. In official statistics the data integration procedures are becoming extremely important due to many reasons: some of the most crucial ones are the cut of the cost, the reduction of response burden and the statistical use of information derived from administrative data. The many possible applications of record linkage techniques and their wide use make them a powerful instrument. The most widespread utilizations of record linkage procedures are the elimination of duplicates within a data frame, the study of the relationship among variables reported in different sources, the creation of sampling lists, the check of the confidentiality of public-use

N. Cibella (✉) · T. Tuoto
Istituto Nazionale di Statistica (ISTAT), via Cesare Balbo 16, 00184 Roma, Italy
e-mail: cibella@istat.it

A. Di Ciaccio et al. (eds.), *Advanced Statistical Methods for the Analysis of Large Data-Sets*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-642-21037-2_8, © Springer-Verlag Berlin Heidelberg 2012

micro-data, the calculation of the total amount of a population by means of capture-recapture models, etc.

Generally, the difficulties in record linkage project are related to the number of records to be linked. Actually, in a record linkage process, all candidate pairs belonging to the cross product of the two considered files, say A and B, must be classified as matches, un-matches and possible matches. This approach is computationally prohibitive when the two data frames become large: as a matter of fact, while the number of possible matches increases linearly, the computational problem raises quadratically and the complexity is $\boldsymbol{O}(n^2)$ (Christen and Goiser 2005). To reduce this complexity, which is an obvious cause of problems for large data sets, it is necessary to decrease the number of comparisons. Then expensive and sophisticate record linkage decision model can be applied only within the reduced search space and computational costs are significantly saved. In order to reduce the candidate pairs space, several methods exist, i.e. techniques of sorting, filtering, clustering and indexing may all be used to reduce the search space of candidate pairs. The selection of the suitable reduction method is a delicate step for the overall linkage procedure because the same method can yield opposite results against different applications.

The debate concerning the performances of different blocking methods is very vivacious among the information technology scientists (Baxter et al. 2003, Jin et al. 2003). In this paper the focus is instead on the statistical advantages of using data reduction methods in performing a probabilistic record linkage process on large data-sets. The outline of the paper is as follow: in Sect. 2 details on the most widespread blocking methods, i.e. standard blocking and sorted neighbourhood, are given; Sect. 3 stresses the statistical point of view on the choice between the compared methods; Sect. 4 reports experimental results proving the statements given in Sect. 3 by means of real data application; finally in Sect. 5 some concluding remarks and future works are sketched.

2 Blocking Methods

Actually, two of the most challenging problems in record linkage are the computational complexity and the linkage quality. Since an exhaustive comparison between all records is unfeasible, efficient blocking methods can be applied in order to greatly reduce the number of pairs comparisons to be performed, achieving significant performance speed-ups. In fact, blocking methods, directly or indirectly, affect the linkage accuracy:

- they can cause missing true matches: when record pairs of true matches are not in the same block, they will not be compared and can never be matched
- thanks to a better reduction of the search space, more suitable, intensive and expensive models can be employed.

So, blocking procedures have two main goals that represent a trade-off. First, the number of candidate pairs generated by the procedures should be small to minimize the number of comparisons in the further record linkage steps. Second, the candidate set should not leave out any possible true matches, since only record pairs in the candidate set will be examined in detail.

The developments in the modern computer power, the machine learning, the data mining and statistical studies improve undoubtedly the performances and the accuracy of the record linkage procedure and help in finding more efficient blocking methods (e.g. the new method with the use of clustering algorithms or high-dimensional indexing). Nevertheless the potential advantages and disadvantages of the several different existing blocking methods make the choice among them a difficult task and there is not a general rule for privileging a method against the others.

In the following subsections, some details on the most widespread blocking methods, i.e. standard blocking and sorted neighbourhood, are given so as to stress the basic characteristics of the two methods compared herewith from a statistical perspective.

2.1 The Standard Blocking Method

The standard blocking method consists of partitioning the two datasets A and B into mutually exclusive blocks where they share the identical value of the blocking key (Jaro 1989) and of considering as candidate pairs only records within each block. A blocking key can be composed by a single record attribute, common to the data sets, or combining more than one attribute. There is a cost-benefit trade-off to be considered in choosing the blocking keys: from one hand, if the resulting blocks contain a large number of records, then more candidate pairs than necessary will be generated, with an inefficient large number of comparisons. From the other hand, if the blocks are too small then true record pairs may be lost, reducing the linkage accuracy. Moreover, to achieve good linkage quality, also the error characteristics of the blocking key is relevant, i.e. it is preferable to use the least error-prone attributes available.

In theory, when the size of the two data sets to be linked is of n records each and the blocking method creates b blocks (all of the same size with n/b records), the resulting number of record pair comparisons is $\boldsymbol{O}(n^2/b)$. This is an ideal case, hardly ever achievable with real data, where the number of record pair comparisons will be dominated by the largest block. So the selection of the blocking keys is one of the crucial point for improving the accuracy of the whole process. To mitigate also the effects of errors in blocking keys, multiple keys can be used and several passes, with different keys, can be performed (Hernandez and Stolfo 1998). Multiple passes improve linkage accuracy but the implementation is often inefficient. Roughly speaking, the multi-pass approach generates candidate pairs using different attributes and methods across independent runs. Intuitively, different runs cover different true matches, so the union should cover most of the true matches. Of course, the effectiveness of a multi-pass approach depends on which attributes are chosen and on the methods used.

2.2 The Sorted Neighbourhood Method

Another of the most well-known blocking method is the sorted neighbourhood one (Hernandez and Stolfo 1995). This method sorts together the two record sets, A and B, by the selected blocking variable. Only records within a window of a fixed dimension, w, are paired and included in the candidate record pair list. The window slides on the two ordered record sets and its use limits to $(2w - 1)$ the number of possible record pair comparisons for each record in the window. Actually, in order to identify matches, the first unit of the list is compared to all the others in the window w and then the window slides down by one unit until the end (Yan et al. 2007). Assuming two data sets of n records each, with the sorted neighbourhood method, the total number of record comparisons is $\boldsymbol{O}(wn)$.

The original sorted neighbourhood method expects a lexicographic ordering of the two data sets. Anyway, records with similar values might not appear close to each other when considering lexicographic order. In general, the effectiveness of this approach is based on the expectation that if two records are duplicates, they will appear lexicographically close to each other in the sorted list based on at least one key.

Similar to standard blocking method whereas the sliding window works as a blocking key, it is preferable to do several passes with different sorting keys and a smaller window size than only one pass with a large window size. Even if multiple keys are chosen, the effectiveness of the method is still susceptible to deterministic data-entry errors, e.g., the first character of a key attribute is always erroneous.

3 A Statistical Perspective in Comparing Blocking Methods

As stated in Sect. 1, when managing huge amount of data, the search space reduction is useful to limit the execution time and the used memory space by means of a suitable partition of the whole candidate pairs space, corresponding to the cross product of the input files. The information technologist community is really active in analyzing characteristics and performances of the most widespread blocking methods as well as in data linkage project at all: a proof of the statement is given by the proliferation of different names to refer the record linkage problem – citation matching, identity uncertainty, merge-purge, entity resolution, authority control, approximate string join, etc. A further evidence is the emergence of numerous organizations (e.g., Trillium, FirstLogic, Vality, DataFlux) that are developing specialized domain-specific record-linkage tools devoted to a variety of data-analysis applications.

In the last years, new attractive techniques for blocking have been proposed: clustering algorithms, high-dimensional indexing, by-gram indexing, canopy. Moreover machine learning methods have been developed in order to define the best blocking strategy for a given problem using training data. Generally speaking, the blocking strategy states a set of parameters for the search space reduction phase: the blocking keys, the method that combines the variables (e.g. conjunction, disjunction), the choice of the blocking algorithms, the window size, the choice of the similarity functions and so on.

In this paper we approach the problem of stating the most suitable blocking method keeping in mind also the statistical perspective on the record linkage problem. In fact, when probabilistic approach is applied, "statistical" problems arise in dealing with huge amount of data. Usually, the probabilistic model estimates the conditional probabilities of being match or un-match assuming that the whole set of candidate pairs is a mixture of the two unknown distributions: the true links and the true non-links (Armstrong and Mayda 1993, Larsen and Rubin 2001). Generally, an EM algorithm is applied in order to estimate the conditional probabilities in presence of latent classification. The statistical problem arises when the number of expected links is extremely small with respect to the whole set of candidate pairs; in other words, if one of the two unknown populations (the matches) is really too small, it is possible that the estimation mechanism is not able to correctly identify the linkage probabilities: it could happen that the EM algorithm still converges, but in fact it estimates another latent phenomenon different from the linkage one. This is why some authors suggest that, when the conditional probabilities are estimated via the EM algorithm, it is appropriate that the expected number of links is not below 5% of the overall compared pairs (Yancey 2004). A solution to this situation is the creation of suitable groups of the whole set of pairs, i.e. a blocking scheme, so that, in each sub-group, the number of expected links is suitable with respect to the number of candidate pairs.

From the statistical perspective, the choice among blocking methods depends on several characteristics, only partially connected to the computational aspects. In this work, some of the most important issues of the blocking strategy are stressed, i.e. the expected match rate and the frequency distribution of the available blocking keys dramatically influence the effectiveness of the chosen blocking method. For instance, if the standard blocking method is really useful to solve linkage problems where the overlap between files is very high, i.e. in de-duplication or post-enumeration survey context, it could be unhelpful when the expected number of matches is very small with respect to the largest file to be linked. Moreover, when most of the blocking key categories are very sparse with low frequencies (no more than five), even if identification power of the key is high, the standard blocking method can't help in defining a probabilistic linkage strategy.

4 Experimental Results

The previous aspects are highlighted in the study of the fecundity of married foreign-women with residence in Italy. This study requires the integration of two data sources: the list of the marriages and the register of births. The two data sets have a common identifier, the fiscal code of the bride/mother. Unfortunately it is affected by errors, particularly when the bride/mother is foreign. Moreover, considering a certain year, the number of births is quite small with respect to the amount of marriages of the same year, so the expected match rate is very low, below the 5% of the largest file.

The data considered in this paper referred to marriages with almost one of the married couple foreign and resident in Lombardy in 2005 and to babies born in the same Region in 2005–2006. The size of each file is about 30,000 records. The common variables are: fiscal code of the bride/mother, the three-digit-standardized name and surname of both spouses/parents, the day/month/year of birth of the bridegroom/father and of the bride/mother, the municipality of the event (marriage/birth). A probabilistic procedure based on EM solution of the Fellegi–Sunter model has been applied.

Due to the file size, a reduction method is needed, avoiding to deal with 900 millions of candidate pairs. The performances of the standard blocking method and of the sorted neighbourhood one are compared.

A previous analysis of the accuracy and of the frequency distribution of the available variables has limited the choice to the three-digit-standardized name and surname of the bride/mother as blocking keys.

We have experimented several strategies in reducing the number of the candidate pairs. The results of the different tests have been compared by means of two different groups of diagnostic for blocking methods: the first one is common in the information technology context while the second one is typical in the statistic community.

The computer scientists currently adopt the reduction ratio (RR) and the pairs completeness (PC) indexes to compare blocking techniques. The RR quantifies how well the blocking method minimizes the number of candidates: $\mathrm{RR} = 1 - C/N$, where C is the number of candidate matches and N is the size of the cross product between data sets. The PC measures the coverage of true matches with respect to the adopted blocking method, i.e. how many of the true matches are in the candidate set versus those in the whole set: $\mathrm{PC} = \mathrm{Cm}/\mathrm{Nm}$ where Cm is the number of true matches in the candidate set and Nm is the number of matches in the whole set. A blocking scheme that optimizes both PC and RR reduces the computational costs for record linkage, decreasing the candidate pairs, and, at the same time, saves the linkage accuracy by means of not loosing true matches. From the statistical perspective, the match rate (MR) is one of the measure, with the linkage error rates, to evaluate the linkage procedure. The MR represents the coverage of true matches of the overall linkage procedure, considering also the applied classification model: MR $= M/\mathrm{Nm}$ where M is the number of true matches identified at the end of the linkage procedure and Nm as already defined.

All these metrics require that the true linkage status for the record pairs is known; we consider as a benchmark the total amount of pairs with common fiscal code and we also refer to such a number when evaluating the improvements in the results of the probabilistic linkage procedures at all. In this study the pairs with common fiscal code are 517 records. As such a key is not error-free, it is possible to find a higher number of pairs that are true matches almost surely, given that they share the same values for high-identification powerful variables: standardized name and surname, day/month/year of birth. This point implies values greater than 1 for the PC and MR metrics.

The first blocking strategy consists in standard blocking method with 3-digit-standardized surname of the bride/mother as key: the categories in each file are about 4,000, resulting in 2,200 blocks and about 580,000 candidate pairs. A probabilistic procedure based on the Fellegi–Sunter model has been applied, considering as matching variables: the three-digit-standardized name of the mother and her day/month/year of birth. The results in terms of matches, possible matches, true matches and MR are shown in Table 1. The relative PC and RR are reported in Table 2.

The inefficiency of standard blocking method, compared to the benchmark, has lead us to test an alternative blocking strategy, based on sorted neighbourhood method. The six-digit-string key composed by joining standardized name and surname and a window of size 15 creates about 400,000 candidate pairs. The probabilistic procedure based on the same Fellegi–Sunter model has been applied and 567 matches and 457 possible matches have been identified. Tables 1 and 2 report the results in terms of matches, possible matches, true matches and MR and PC and RR respectively.

The standard blocking method was tested also with six-digit-string name and surname but, due to the about 24,000 categories of this key, often without any overlap between the two files, the EM algorithm for the estimation of the probabilistic

Table 1 Statistical diagnostics for blocking strategies comparison

	Blocking Surname three-digit	Sorted neighbourhood surname six-digit
Matches	439	567
Possible matches	359	457
True matches	448	592
MR	0.867	1.145

Table 2 Computer scientist diagnostics for blocking strategies comparison

	Blocking surname three-digit	Sorted neighbourhood surname six-digit
PC	1.064	1.145
RR	0.999	0.999

linkage parameters doesn't work so the MR is not evaluable. Anyway, the PC and RR are equal to 1.039 and 0.999.

As showed in the above tables, the differences between the two blocking strategies emerge basically from the statistical perspective, see the MR values, whereas the measures in Table 2 highlight smoothed differences.

5 Concluding Remarks and Future Works

The linkages presented in this paper have been performed by RELAIS, an open source software designed and implemented by ISTAT. It provides a set of standard methods and techniques in order to execute record linkage applications. In order to better face with the complexity of linkage problem, it is decomposed into its constituting phases; the software allows the dynamic selection of the most appropriate technique for each phase and the combination of the selected techniques so that the resulting workflow is actually built on the basis of application and data specific requirements. In fact, RELAIS has been designed with a modular structure: the modules implement distinct record linkage techniques and each one has a well defined interface towards other modules. In this way it is possible to have a parallel development of the different modules, and to easily include new ones in the system. Moreover, the overall record linkage process can be designed according to specific application requirements, combining the available modules. The RELAIS approach overcomes the question on which method is better compared to the others, being convinced that at the moment there is not a unique technique dominating all the others. The strength of RELAIS consists in fact of considering alternative techniques for the different phases composing the record linkage process. RELAIS wants to help and guide users in defining their specific linkage strategy, supporting the practitioner's skill, due to the fact that most of the available techniques are inherently complex, thus requiring not trivial knowledge in order to be appropriately combined. RELAIS is proposed also as a toolkit for researchers: in fact, it gives the possibility to experiment alternative criteria and parameters in the same application scenario, that's really important from the analyst's point of view. For all these reasons RELAIS is configured as an open source project, released under the EU Public License.

This paper is a first step in comparing blocking methods for record linkage, keeping in mind the statistical perspectives. Further tests are needed. It could be useful for instance to exploit data sets where the true linkage status is completely known; unfortunately these is hard to achieve without a clerical review, but manual checks are quite prohibitive for very large data-sets. A possible approach to these issues could be to replicate these experiments with synthetic data sets.

Further analyses can also be useful in comparing blocking methods in other contexts, with an expected match rate intermediate compared with the post-enumeration survey one, that is about the 99%, and with that considered in this paper, that is lower than the 5% of the largest file.

Other goals of future studies concern the examine of the statistical impact of more complicated blocking methods, such as the bigram indexing, the canopy clustering, etc. and the evaluation of the comparability of blocking choices suggested by domain experts with respect to those learned by machine learning algorithms, both supervised and unsupervised and fully automatic ones.

References

Armstrong J.B. and Mayda J.E (1993) Model-based estimation of record linkage error rates. S*urvey Methodology*, 19, 137–147

Baxter R., Christen P., Churches T. (2003) A comparison of fast blocking methods for record linkage http://www.act.cmis.csiro.au/rohanb/PAPERS/kdd03clean.pdf

Christen P. and Goiser K. (2005) Assessing duplication and data linkage quality: what to measure?, *Proceedings of the fourth Australasian Data Mining Conference,* Sydney, December 2005, http://datamining.anu.edu.au/linkage.html

Hernandez M.A. and Stolfo S.J. (1995) The merge/purge problem for large databases. In M. J. Carey and D. A. Schneider, editors, *SIGMOD*, pp. 127–138

Hernandez M.A. and Stolfo S.J. (1998) Real-world data is dirty: Data cleansing and the merge/purge problem. *Data Mining and Knowledge Discovery* 2(1), 9–37

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

Jin L., Li C., Mehrotra S. (2003) Efficient Record Linkage in Large Data Sets. *Proceedings of the 8th International Conference on Database Systems for Advanced Applications* (DASFAA) 2003, Kyoto, Japan, http://flamingo.ics.uci.edu/pub/dasfaa03.pdf

Larsen M.D. and Rubin D.B. (2001). Iterative automated record linkage using mixture models. *Journal of the American Statistical Association*, 96, 32–41

Relais, Record linkage at Istat, http://www.istat.it/it/strumenti/metodi-e-software/software/relais

Yan S., Lee D., Kan M.-Y., Giles C. L. (2007) Adaptive sorted neighborhood methods for efficient record linkage, JCDL'07, Vancouver, British Columbia, Canada

Yancey W.E. (2004) A program for large-scale record linkage. In Proceedings of the Section on Survey Research Methods, *Journal of the American Statistical Association*