

Statistical perspective on blocking methods when linking large data-sets

Nicoletta Cibella, Tiziana Tuoto

Istituto Nazionale di Statistica (ISTAT), via Cesare Balbo 16, 00184 Roma, Italy,
{surname}@istat.it

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.

Keywords: record linkage, blocking, sorted neighborhood

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, the cut of the cost, the reduction of response burden and the use of information derived from administrative data are some of the most crucial ones.

There many possible applications of such techniques and their wide use makes them a powerful instruments. Just mentioning the most widespread record linkage procedures utilization: the elimination of duplicates within a data frame; the study of the relationship among variables reported in different sources; the creation of a sampling list; the check of the confidentiality of public-use microdata; the calculation of the total amount of a population by means of capture-recapture models; etc. Generally, the difficulty is related to the number of records to be linked so as to identify the same real world entity. Actually, in a record linkage process, all candidate pairs, those belonging to the cross product of the two files considered, say A and B , must be classify as matches, un-matches and possible matches. When dealing with large datasets, the comparison of the matching variables is almost impracticable; as a matter of fact, while the number of possible matches increases linearly, the computational problem raises quadratically, the complexity is $O(n^2)$ (Christen and Goiser, 2005). To reduce this complexity, which is an obvious cause of problems for large databases, it is necessary to decrease the number of pairs, say a, b (with a belongs to A and b belongs to B), to be compared. Starting from this reduced search space, different decision models can be applied to determine whether a pair of records ($a; b$) is a match, an un-match or a possible match. Techniques of sorting, filtering, clustering and indexing may all be used

to reduce the search space of candidate pairs. The debate concerning the performances of different blocking methods is very vivacious among the information technology scientists (Baxter 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.

2. Blocking methods

Actually, the major challenges in record linkage are the computational complexity and the linkage accuracy: efficient blocking methods can 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 comparisons ratio, more suitable, intensive and expensive solutions comparators can be employed.

In this section, some details on the most widespread blocking methods, i.e. standard blocking and sorted neighbourhood, are given.

The standard blocking method consists of partitioning the two datasets into blocks where they share the identical value of the blocking key (Jaro, 1989) and of considering linkable 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 records pairs than necessary will be generated, leading to an inefficiently 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 accuracy, also the error characteristics of the blocking key is relevant, i.e. it is preferable to use the least error-prone attributes available. To mitigate the effects of errors in blocking keys, multiple keys can be used and several passes, with different keys, can be performed. Multiple passes improve linkage accuracy but the implementation is often quite inefficient.

In theory, when the size of the two data sets to be linked is of n records each and the blocking methods created b blocks (all of the same size with n/b records) the resulting number of record pair comparisons is $O(n^2/b)$. 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.

Another of the most widespread blocking method is the sorted neighbourhood one. This method sorts the two record sets by the same variable. Only records within a window of a fixed dimension, w , are paired with each other and included in the candidate record pair list. The window slides on the two ordered record sets and its use limits the number of possible record pair comparisons for each record in w to $(2w-1)$. Assuming two data sets of n records each, with the sorted neighbourhood method, the total number of record comparisons is $O(wn)$.

Similar to standard blocking method, it is advantageous to do several passes with different sorting keys and a smaller window size, than only one pass with a large window size.

With this method one problem arises if a number of record larger than the size of the fixed window have the same value in the sorting key.

3. A statistical perspective in comparing blocking methods

The search space reduction can be useful for two different reasons: as above described, when managing huge amount of data, it is useful to reduce the execution time and the used memory space by means of a suitable partition of the whole candidate pairs space, coming from the cross product of the input files. Moreover, the record linkage probabilistic approach brings statistical problems in dealing with huge amount of data. In fact, the probabilistic models, generally used in order to estimate the conditional probabilities of being link or being non-link, do not allow to correctly identify such probabilities when the number of possible links is extremely small with respect to the whole set of candidate pairs. In particular, 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). Also the statistical problem can be overcome by means of the creation of suitable groups or partitions of the whole cross product set of pairs, 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 are stressed: the expected match rate and the frequency distribution of the available blocking keys dramatically influence the efficacy 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 and concluding remarks

The previous aspects are highlighted in the study of the fecundity of married foreign-women with residence in Italy. In this context, the number of birth rate is quite small with respect to the amount of marriages, so the expected match rate is very low (below the 5% of the largest file).

The data of the marriages with almost one of the two spouse foreign, with the residence in Lombardy in 2005, has been linked with the registers of births in the same Region in the two-year period following the marriage, exclusively via a probabilistic approach. The size of each file is about 30 000 records. The variables, common to the two files, are: fiscal code of the bride/mother, the 3-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 data size, a data reduction method is needed; the paper compares the performances of the above-mentioned two blocking methods. An analysis of the accuracy and of the frequency distribution of the available variables has limited the choice to the 3-digit-standardized name and surname of the bride/mother as blocking keys; the day/month/year of birth are not useful because of the huge amount of records in each of the derived blocks.

We have experimented some different strategies in reducing the number of the candidate pairs. First of all, due to the high quality of the standardized name and surname, a blocking key has been composed from joining name and surname, obtaining a 6-digit-string blocking key. The standard blocking method, based on this key, doesn't work, due to the 24 000 categories of this key, often without overlap between the two files, that produces 1 400 blocks, with very low pairs. Secondly, we have considered only the 3-digit-standardized surname of the bride/mother: in this case, the categories are about 4 000, producing about 2 200 blocks. A probabilistic procedure based on the Fellegi-Sunter model has been applied, considering as matching variables: the 3-digit-standardized name of the mother and her day/month/year of birth. The linkage produces 439 matches and 359 possible matches. Among all the matches, only 109 are surely true matches, because of the concordance in the fiscal code or in all the variables referred to the bridegroom/father (the 3-digit-standardized name and surname and the day/month/year of birth), while among possible matches, only 9 have the same fiscal code or a concordance in all the bridegroom/father variables.

The inefficiency of these results has lead us to change blocking strategy: we applied an alternative procedure based on sorted neighbourhood method. The 6-digit-string of name and surname over a 15 window size creates about 400 000 candidate pairs, on which the previous Fellegi-Sunter model has been applied. The procedure identified 567 matches and 457 possible matches. Among the matches, even 499 pairs have the same fiscal code or agree in all the bridegroom/father variables while, among the possible matches, the concordance in the pairs are 25; so, totally, more than 590 true matches are identified by this procedure, while the total amount of pairs with common fiscal code in the files are 517 records.

The linkage is performed by RELAIS, an open source software developed by ISTAT, that provides a set of standard methods and techniques in order to execute record linkage applications.

References

- 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>
- 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*, Volume 84, pp. 414-420.
- 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*.