Multi-Source Object Identification With Constraints

(Research-in-Progress)

Matteo Di Gioia
DIS-Università di Roma "Sapienza"
digioia@dis.uniroma1.it

Domenico Beneventano
DII-Università di Modena e Reggio Emilia
domenico.beneventano@unimore.it

Monica Scannapieco
Istat - Istituto Nazionale di Statistica
scannapi@istat.it

Abstract: The problem of identifying the manifold generated copies of an object is known as Object Identification (OI). Numerous solutions have been proposed to solve this task, based on the similarity between two objects. Most of these solutions are oriented to discover *pairs* of duplicates (pairs-oriented OI) rather than *sets* of similar objects (group-oriented OI), for which some clustering techniques are used. In this paper, we proposed a new technique, based on the concept of constraints, to resolve the group-oriented OI problem. It is composed of two phases: extraction phase and grouping. During the extraction phase constraints are extracted by analyzing data at hand. After that we have collected the constraints, we reason about those to find the groups of similar objects. The group-based OI technique we propose allows us to deal with multiple sources.

Key Words: Record Linkage, Entity Resolution, Constraints, Multi-source Object Identification, Clustering

1. Multi-Source Object Identification: introduction and proposal

The OI problem is a real problem in data integration; in fact the integration of different sources introduces the presence of multiple identifiers for the same entity [2]. Most of the research activity was focused, until now, on the resolution of duplicate detection or on the case in which two sources are matched against each other (n=2). These solutions are oriented to discover *pairs* of duplicates (pairs-oriented OI) rather than *sets* of similar objects (group-oriented OI). In this article, we deal with the resolution of the group-oriented problem in a multi-source context. Some works on this problem include the Rendle's approach [4] and the Bhattacharya's one [1]. In our work, we focus on a `1:n` matching, i.e. there are n copies for a real entity.

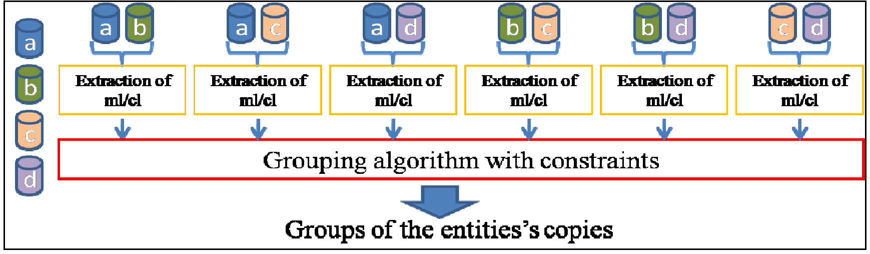

Figure 1: Example of execution of the new object identification process.

In particular, we extend the work of Rendle [4] analyzing the limitations of the use of pairwise techniques in a multi-source environment and identifying some properties that must be guaranteed in order that the pairwise techniques could be suitable to the multi-source environment. We also describe a technique to perform multi-source object identification in a "safe" mode, i.e. respecting some properties.

Many times the multi-source problem is resolved reducing all sources to one and then applying a classic method on the new unified sources [3]. This approach does not seem to be appropriate for two main reasons: first, the necessary space grows with the sources' number; second, the process is not oriented to the only duplicates but always elaborates the whole set of data sources, i.e. in the different steps all the objects are verified also the ones without duplicates. In our work, we focus the attention on the OI problem excluding the objects without duplicates; if in a process step an object is considered without duplicates, then it is no longer verified during the subsequent steps. To do it, first we search some constraints between objects and then we use the ones found to discover the clusters of objects. In particular, instead of starting from clusters composed of only one element, we can start from clusters disguised as constraints (derived by human or computed knowledge).

The cluster object identification process between more than two sources that we propose is composed of two stages, showed in fig.1 (see [5] for a deeper description). In a first stage, the *n* sources are analyzed to extract some useful information, while during the second stage the results of the first stage are collected and analyzed to perform the grouping of the duplicates. During the first stage, the sources are analyzed in pairs to extract relations of similarity or not. With more precision, during this phase two types of relation are extracted, *must links* and *cannot links*, which correspond to similarity links between objects. We use the SNM [3] method to find the ML and CL pairs. Notice that we use also the negative information provided as output of the SNM. In the second stage, the grouping phase, the pairs extracted during the first stage are elaborated to form the groups of the entities' copies. The objects of the *ml* pairs are grouped together if the pairs have some extremity in common (i.e. one object in common) and if after the grouping they continue to respect the set of relations extracted (the extracted relations become constraints in the second phase). We outline that the sets of *ml* and *cl*, generated by the extraction phase, can be overlapping. This is due to the possibility to execute the SNM (or other techniques) more than one time with different settings. To generate the groups of objects we use our algorithm that starts with a cluster composed by two elements which belongs to a relation and then extends this cluster until is possible. During the expansion of the cluster when a *ml* is added to expand the cluster, it is verified that its addition doesn't produce any conflicts with the *cl* constraints. If there aren't conflicts then the *ml* pair is added to the cluster. If there are some conflicts then the combination that guarantees the best clusterization must be chosen. We have studied and evaluated our technique which exhibits good effectiveness performance (due to lack of space we don't quote the evaluation and algorithm details).

In conclusion, in our work, we have considered the problem of the object identification among many sources. We have posed the attention to the distributed environment and thus we have searched a solution adaptable to a distributed environment. For this reason, we considered the concept of the constraints, in the meaning of our work, as a promising step to distributed object identification.

2. REFERENCES

[1] I. Bhattacharya and L. Getoor. *Collective entity resolution in relational data.* ACM Trans. Knowl. Discov. Data, 1(1):5, 2007.

[2] I. P. Fellegi and A. B. Sunter. *A theory for record linkage.* Journal of the American Statistical Association, 64(328):1183–1210, 1969.

[3] M. A. Hernandez and S. J. Stolfo. *The merge/purge problem for large databases.* In In Proceedings of the 1995 ACM SIGMOD, pages 127–138, 1995.

[4] S. Rendle and L. Schmidt-Thieme. *Object identification with constraints.* Data Mining, 2006. ICDM '06. Sixth International Conference on, pages 1026–1031, Dec. 2006.

[5] Technical report on site: http://www.dis.uniroma1.it/~digioia/documents/MSOIWC.pdf