Chapter 9
Overview of MERA: An Architecture to Perform Record Linkage in Music–Related Databases

Daniel Fernández-Álvarez
University of Oviedo, Spain

José Emilio Labra Gayo
University of Oviedo, Spain

Daniel Gayo-Avello
University of Oviedo, Spain

Patricia Ordoñez de Pablos
University of Oviedo, Spain

ABSTRACT

The proliferation of large databases with potentially repeated entities across the World Wide Web drives into a generalized interest to find methods to detect duplicated entries. The heterogeneity of the data cause that generalist approaches may produce a poor performance in scenarios with distinguishing features. In this paper, we analyze the particularities of music related-databases and we describe Musical Entities Reconciliation Architecture (MERA). MERA consists of an architecture to match entries of two sources, allowing the use of extra support sources to improve the results. It makes use of semantic web technologies and it is able to adapt the matching process to the nature of each field in each database. We have implemented a prototype of MERA and compared it with a well-known music-specialized search engine. Our prototype outperforms the selected baseline in terms of accuracy.

DOI: 10.4018/978-1-5225-7186-5.ch009
INTRODUCTION

Although the problem of entity reconciliation has been largely studied, it remains a challenging issue. New research trends related to entity reconciliation has appeared in the last decade. This includes the need of developing efficient algorithms to deal with Big Data (Castano, Ferrara, & Montanelli, 2018; Enríquez, Domínguez-Mayo, Escalona, Ross, & Staples, 2017), the challenge of linking individuals in domains in which preserving their privacy is a requirement (Pow et al., 2017) or the need to align ontologies in Linked Data scenarios (Achichi et al., 2016; Zahaf & Malki, 2018).

Different entity reconciliation environments rise different challenges. The specific context of music databases has not been deeply studied, despite the content related to this kind of datasets include a set of insightful features. Examples of fields usually contained in musical databases are titles, artist names, albums, genres, etc. Each one of these fields have some distinguishing peculiarities which cause that a certain real entity may be represented in different ways in different databases. For instance, there are many specific correct forms, or at least recognizable forms, in which we could express the name of an artist. This includes artistic names, civil names, names conventions (“The Beatles” vs “Beatles, the”), acronyms or common misspellings. When dealing with information related to genre, one may find that a certain song is specified as pop in a database, as rock in a second one and as pop-rock in a third one. Sometimes, the same genre is even named with different forms that are in fact expressing the same reality.

Our assumption is that finding general reconciliation rules between two databases is far from being trivial, as well as finding appropriate rules or strategies to conciliate each field of those databases. The result could drastically change if it is compared to the rules that may be used when handling a different pair of sources. Trying to establish general rules could drive into an unnecessary number of failures when identifying two records of different databases as forms of the same real entity. The inference of reconciliation rules in a particular case through the use of training data may be useful for covering issues such as misspellings, naming conventions or even noisy prefixes/suffixes, but it cannot handle cases in which the strings that represent the entities do not have common characters (example: “The King of Rock” should be recognized as “Elvis Presley”).

Our main contribution is the specification of MERA architecture. MERA tries to adapt to all those scenarios using graph concepts and semantic web technologies. Our approach turns the information of one of the target databases into a custom RDF graph $G$ containing all the information (name variations, alias, common misspellings ...) of every database record, as well as the relations between those records. The records of the second database are turned into complex queries that will be launched against $G$. The result of each query is the list of the most similar nodes to the target record according to:

- String-distance-based functions.
- Use of all the alternative identifying forms of a concept.
- Graph navigation in order to detect shared associated entities for disambiguation purposes.

MERA can use different reconciliation algorithms for each pair of databases and even for each field of those databases, trying to cover all the issues linked to the nature of the data. Our solution is able to reach better results with more prior knowledge of the data issues, since the user is the agent that specifies the algorithms to use. MERA allows configuring different properties that should be considered, the reconciliation algorithms to apply in each case, and the threshold of similarity that a result must reach to be accepted. It also provides mechanisms to incorporate ad-hoc algorithms in the reconciliation process.