MERA: A Musical Entities Reconciliation Architecture Based on Semantic Technologies

Daniel Fernández-Álvarez, Department of Computer Science, University of Oviedo, Oviedo, Spain
Jose Emilio Labra Gayo, University of Oviedo, Oviedo, Spain
Daniel Gayo-Avello, Department of Computer Science, University of Oviedo, Oviedo, Spain
Patricia Ordóñez de Pablos, Department of Business Administration. Faculty of Economics and Business, University of Oviedo, Oviedo, Spain

ABSTRACT

In this paper, the authors describe Musical Entities Reconciliation Architecture (MERA), an architecture designed to link music-related databases adapting the reconciliation techniques to each particular case. MERA includes mechanisms to manage third party sources to improve the results and it makes use of semantic technologies, storing and organizing the information in RDF graphs. They have implemented a prototype of their approach and have used it to link sources with different levels of data quality. The prototype has been effective in more than 94% of the cases under the conditions of their experiments. The authors have also compared their prototype with a well-known music-specialized search engine, outperforming the search results in the two experiments that they performed.

KEYWORDS

Adaptable Architecture, Collective Matching, Entity Reconciliation, Music Metadata, RDF Graph, Record Linkage, Semantic Search

1. INTRODUCTION

Although the problem of entity reconciliation has been largely studied, it remains a challenging issue. The proliferation of large databases with potentially repeated entities across the World Wide Web drives into an interest to find methods to detect duplicated entries when no reliable unique identifiers are available. In this paper, we provide an architecture for the specific task of linking records of musical databases. The Musical Entities Reconciliation Architecture (MERA) discovers links between elements of different databases that represent the same real-world entity in the music domain. Our approach is able to adapt the linkage process to the different content and nature of each database, letting the user configure different reconciliation algorithms for different attributes or type of entities.

Examples of fields usually contained in musical databases are titles, artist names, albums, genres, etc. The task of recognizing these kinds of contents is strongly connected to the record linkage problem, since it consists of the detection of records or entries referring to the same real-world entity. However, we have designed MERA with the assumption that the type of metadata linked to the music world presents a certain number of peculiarities that should be considered. For instance, there are many

DOI: 10.4018/IJSWIS.2017100103
specific cases of correct forms, or at least recognizable forms, in which we could express the name of an artist, including but not limited to:

- Artistic names vs civil names, e.g., “Stefani Joanne Angelina Germanotta” or “Lady Gaga”;
- Naming conventions, e.g., “The Beatles” or “Beatles, the”;
- Official or widely extended alias, e.g., “The King of Rock” instead of “Elvis Presley”;
- Mixings between civil names and artistic names, e.g., “Shakira”, “Shakira Isabel Mebarack Ripoll”, “Shakira Mebarack”, etc.;
- Acronyms, e.g., “System of a down” or “SOAD”;
- Usual misspellings, e.g., “Bruce Springsting” instead of “Bruce Springsteen”;
- Name of an artist linked to a song that should actually be linked to a group, e.g., “Michael Jackson” instead of “The Jackson Five”.

Issues such as misspellings or acronyms are not specific of music-related metadata, and they can be found in databases or datasets of different nature. However, issues such as the existence of both artistic and civil name are exclusive of artists’ specification. By contrast, when trying to conciliate other types of musical entities, e.g., songs, a different set of specific problems related to the nature of songs may appear. An example could be the management of the word “feat” (or variations such as “ft.”, “featuring”, etc.). When “feat” appears in a song title, it usually means that in that title there is a name of a collaborator included. Both “feat” and its following words may be discarded from the song name itself. However, they can possibly be computed in some other way since they can become very useful information.

In noisy or hand-made databases it is also possible to find extra words at the beginning or at the end of a song title. For instance, sequences linked to the radio program in which a song was played or to the place of a live performance. This may become even more troublesome in especially noisy databases such as those formed by the compilation of standalone audio files’ metadata. When handling audio files wrongly labeled, it is possible to find titles that in fact contain all the associated metadata (artist, date, genre...) in a single field.

Another example of a musical concept that presents associated issues due to its special nature is the genre. When dealing with genres, it could happen that the same 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 (false positives/negatives) 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 handy for covering issues such as misspellings, naming conventions or even noisy prefixes/suffixes, but they 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 in this paper 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, etc.) of every database record, as well
Semantic Multimedia Information Analysis for Retrieval Applications
www.igi-global.com/chapter/semantic-multimedia-information-analysis-retrieval/28934?camid=4v1a