

TV Program Guide Metadata Acquisition

Peter DUBEC*

*Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
xdubecpl@is.stuba.sk*

Today we can find huge amounts of metadata for multimedia items published online in different forms. Currently main problem is that items are published online in different languages, and we are trying to map these items on Slovak entities. We see a big problem in how titles of items of TV program guide are translated, because many times some words are translated differently or whole local title is completely different than original title so then it is a problem to find adequate entity published on online sources. Our goal is to focus our work on TV program guide metadata enrichment aimed at mapping entities of Slovak TV program guide on entities published online for the purpose of improving quality of these services and also to enable personalized recommendation of TV program guide items [1].

It is very important to find relevant and high quality metadata to describe items of TV program guide. Thus as a source of metadata we use two big online movie databases, which can be considered reliable, because publishing of metadata is supervised by administrators. First is CSFD where data are published in Czech or Slovak language and also have good structure so it is easy to work with them and we are also using CSFD as a source for our TV program guide. Second database which we used as a source for metadata extraction is IMDB which is international movie database where data are published in English language. [2]. We employ Linked Data [3] as a next source for metadata. But we are extracting metadata only from two datasets Dbpedia [4] and Freebase. As a final source of metadata we take our own automatically generated metadata, obtained from documentary movies classification.

Our method of metadata acquisition consists of three steps:

- TV program guide extraction and basic metadata acquisition
- mapping items on Linked Data space and metadata extraction
- generating new metadata based on documentary movies classification

First step of our method is parsing of TV program guide from CSFD database. During parsing of TV program, when an item is extracted, it is automatically mapped to its

* Supervisor: Mária Bieliková, Institute of Informatics and Software Engineering

representing entity in the CSFD movie database. Second step is mapping extracted entities on its representing entities in the International Movie Database (IMDB).

We also extract metadata from entities in Linked Data space. Process of mapping items on Linked Data space sends specified queries in SPARQL for DBpedia endpoint and in MQL for Freebase endpoint. These queries are constructed from item attributes, which we have available in English and on beginning these queries are very strict, what means they contains as many attributes as possible. When there is no match found, some attributes are deleted from query to make it more general and to increase possibility of finding match.

Final step of our method is enriching TV program guide with our own automatically generated metadata and consists of following steps:

- translate item description into English language
- tag item with automatically generated keywords
- divide items into subcategories based on generated keywords

In our work we enrich Slovak TV program guide with metadata extracted from online movie databases CSFD, IMDB and also from Linked Data space. Moreover, we concentrate on one specific category of TV program items - documentary movies for which we acquire additional metadata based on documentary movies classification. This work is a part of larger project aimed at personalized recommendation of TV program for Slovak viewers. This purpose primarily influenced our focus of metadata enrichment.

Amended version was published in Proc. of the 11th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2015), STU Bratislava, 135-139.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG 1/0646/15.

References

- [1] Hu, Y., Wang, Z., Wu, W., Guo, J., Zhang, M.: Recommendation for Movies and Stars Using YAGO and IMDB. In: Web Conference (APWEB), 2010 12th International Asia-Pacific, Busan, Korea, IEEE, 6-8 April 2010, pp. 123 – 129.
- [2] Peralta, V.: Extraction and integration of movielens and imdb data. Technical report, Tech. Report, Laboratoire PRiSM, Université de Versailles, France, 2007.
- [3] Berners-Lee, T.: Linked Data. Available at <http://www.w3.org/DesignIssues/LinkedData.html>, 2006, [Accessed 01.12.2014].
- [4] Exner, P., Nugues, P.: Entity extraction: From unstructured text to DBpedia RDF triples. In: The Web of Linked Entities Workshop (WoLE 2012), Boston, USA, 2012.