Linking Multimedia to Microblogs for Metadata Extraction

Peter GAŠPAR*

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

With the expansion of the open Web, information overload has become a huge problem. Many information retrieval approaches are trying to deal with retrieving, representation, searching and storing of large datasets. Metadata are widely used to describe instances - especially in the domain of multimedia. They are usually generated from the related content. However, many sources of related content do not provide efficient and complete related content. It is also difficult to find the sources, which we can consider reliable enough.

If we are talking about related content, we should distinguish between the static and the dynamic one. From the user's point of view, a *dynamic related content* might be more attractive. Gathering dynamic related content usually involves processing a *user-generated content*: reviews, ratings or even page-visits. Nowadays one of the most popular sources of dynamic related content is microblogs. Each instance of content is called *post* and it covers for example statuses (or thoughts), pictures, videos and external links.

Many researchers were trying to map the general knowledge base (i.e. Wikipedia) to posts from microblogs. Cremonesi et al. in [1] focused on analysing tweets about the TV content. They mapped tweets using exact word match, free match (similar to n-grams) with TV shows' title, but they also introduced a one-class SVM classifier. Another approach that was combining multiple ideas was proposed by Gattani et al. in [2]. They were using a knowledge base (Wikipedia) as the main source of information for mapping. Additionally, they included the context of social posts and the user's activity to classify posts that did not contain any named entities.

We are introducing a mapping approach, which is using the titles of TV shows from the TV schedule. The main idea is to find the similarity between the titles of TV shows and the posts from the microblog. This similarity is based on the patterns used by the producers of the content in the microblog's posts. We have analysed the social content

* Supervisor: Jakub Šimko, Institute of Informatics, Information Systems and Software Engineering

Spring 2016 PeWe Workshop, April 2, 2016, pp. 53–54.

posted by 11 TV channels (Slovak and worldwide). Based on our analysis we have identified three key features of the posts that are frequently used by the authors of the posts: *named entities*, *hashtags*, *external links*. The process of our mapping consists of these main steps:

1. Normalize TV shows' titles by removing all diacritical marks (e.g. á, é, ů), removing any information about TV series (series number, current episode), and lowercasing all the capital characters.
2. Generate the list of *prefix words* from the TV shows' titles.
3. Generate candidate named entities and hashtags from the titles of TV shows.
4. Extract named entities, hashtags and external links from the microblog post.
5. Find the relationship between the extracted and generated sets using n-gram similarity. Final similarity between the TV show and the microblog post is an average of all particular n-gram similarities.

A *Prefix word* contains up to 4 characters from the first word of the title. We use prefix words in the process of candidate entities generation to eliminate words that are not suitable to be candidates (e.g. the first word in the sentence often starts with the capital letter even if it is not a named entity).

To evaluate our solution, we collected Facebook posts from Slovak TV stations (Markiza, JOJ) and manually labelled 520 of them with the corresponding related TV shows. We have also introduced a threshold parameter rel_{THRES}, which allows us to dismiss mapping results that are supposed to be false positives. For $rel_{THRES} \in [60,70]$ we have scored in average precision 88%, recall 73%, and F1-measure 80%.

In our work we have analysed behaviour of users on microblogs. We have proposed a new method to map shows from the TV schedule to the posts from the microblogs. To evaluate our solution, we have made several experiments. Our results are promising and comparable to related works. In our future work we will try to find a solution for the case, when the microblog post does not contain any named entities, hashtags, nor external links.

Extended version was published in Proc. of the 12th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2016), STU Bratislava.

References

[1] Cremonesi, P., Pagano, R., Pasquali, S., Turrin, R.: TV Program Detection in Tweets. In: *Proceedings of the 11th European Conference on Interactive TV and Video*. EuroITV '13, New York, NY, USA, ACM, 2013, pp. 45-54.

[2] Gattani, A., Lamba, D.S., Garera, N., Tiwari, M., Chai, X., Das, S., Subramaniam, S., Rajaraman, A., Harinarayan, V., Doan, A.: Entity Extraction, Linking, Classification, and Tagging for Social Media: A Wikipedia-based Approach. *Proc. VLDB Endow.*, 2013, vol. 6, no. 11, pp. 1126-1137.