Topical Semantic Recommendations for Auteur Films

Christian Rakow, Andreas Lommatzsch, and Till Plumbaum
Technische Universität Berlin – DAI-Labor
Ernst-Reuter-Platz 7, D-10587 Berlin, Germany
{Christian.Rakow, Andreas.Lommatzsch, Till.Plumbaum}@dai-labor.de

ABSTRACT

With the ubiquity of fast internet connections and the growing availability of Video-On-Demand (VOD) services powerful recommender systems are needed. Traditionally, movie recommender systems apply user-based collaborative filtering providing high quality recommendations if users maintain user profiles describing preferences and movie ratings. The shortcomings of Collaborative Filtering are that comprehensive user profiles are required and users tend to get recommendations very similar to the user profile ("filter bubble"). In addition, CF-based recommenders neither consider current trends nor the context. In order to overcome these weaknesses, we develop a system identifying interesting events in the stream of current news and deploying this information for computing recommendations. Our system gathers topics of interest from TWITTER and RSS-Feeds, extracts relevant Named Entities, and uses semantic relations for recommending movies closely related to these topics. We explain the used algorithms and show that our system provides highly relevant recommendations.

Keywords

movie recommendations; semantic graphs; named entities

1. INTRODUCTION

Movie recommendation systems have been a widely studied topic for many years. The most frequently used approach is Collaborative Filtering (CF) based on the assumption that users who showed similar preferences in the past will like the similar movies in the future. CF-based algorithms deliver high quality results if a sufficient number of ratings for all movies is available and user profiles contain a sufficient number of preferences. Known weaknesses of CF-based approaches are: The bias towards popular movies, the "cold-start problem", and the missing consideration of the user's context as well as currents trends. These shortcomings make CF-based approaches inappropriate for niche markets characterized by limited number of ratings and specific user preferences. In this paper we present a system tailored to the needs of a VOD portal specialized on auteur films. In contrast to the main stream VOD companies, the film catalog focuses on the genres art house, independent, and documentaries. Specific challenges (when developing a recommender for this portal) are the limited user feedback, documentaries (having features differing from regular movies), and the specific interest of the target group.

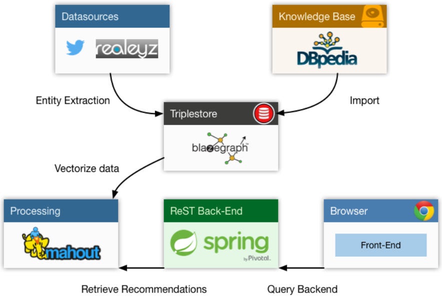


Figure 1: The figure visualizes the main components of our recommender system.

In order to provide highly relevant recommendations and to overcome the shortcomings of CF, we decided to develop a recommender strategy that analyzes news streams (relevant for the target group) and identifies the events suitable for computing topical recommendations. For each event we recognize related Named Entities and compute the films semantically connected with these entities.

We present our system for providing film recommendations based on the analysis of current news and TWITTER messages. We explain the architecture of the system and discuss the used algorithms in detail (Sec. 2). A description of the recommender service and a preliminary evaluation are given in Sec. 3. Finally, Sec. 4 gives a conclusion and an outlook on future work.

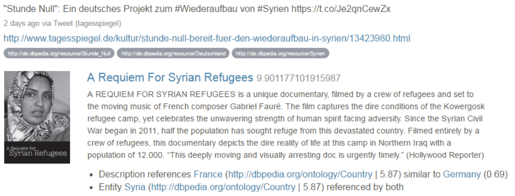
2. SYSTEM OVERVIEW

The system is built of components visualized in Fig. 1. We discuss the components tailored for crawling relevant documents, identifying relevant Named Entities, and computing recommendations in the next Section.

News Retrieval and Semantic Knowledge Bases

We collect news relevant for the target group of the recommender system. For this purpose we consider three different sources: (1) TWITTER messages published by an expert-

RecSys '16 September 15-19, 2016, Boston , MA, USA
© 2016 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-4035-9/16/09.
DOI: http://dx.doi.org/10.1145/2959100.2959110



"Stunde Null": Ein deutsches Projekt zum #Wiederaufbau von #Syrien https://t.co/Je2qnCewZx

2 days ago via Tweet (tagesspiegel)

http://www.tagesspiegel.de/kultur/stunde-null-bereit-fuer-den-wiederaufbau-in-syrien/13423980.html

http://de.dbpedia.org/resource/Stunde_Null
http://de.dbpedia.org/resource/Deutschland
http://de.dbpedia.org/resource/Syrien
A Requiem for Syrian Refugees

A Requiem For Syrian Refugees 9.901177101915987

A REQUIEM FOR SYRIAN REFUGEES is a unique documentary, filmed by a crew of refugees and set to the moving music of French composer Gabriel Fauré. The film captures the dire conditions of the Kowergosk refugee camp, yet celebrates the unwavering strength of human spirit facing adversity. Since the Syrian Civil War began in 2011, half the population has sought refuge from this devastated country. Filmed entirely by a crew of refugees, this documentary depicts the dire reality of life at this camp in Northern Iraq with a population of 12.000. "This deeply moving and visually arresting doc is urgently timely." (Hollywood Reporter)

- Description references France (http://dbpedia.org/ontology/Country | 5.87) similar to Germany (0.69)
- Entity Syria (http://dbpedia.org/ontology/Country | 5.87) referenced by both

Figure 2: The figure shows a example recommendation. Based on the Named Entites recognized in a Twitter message, a movie is recommended. The semantic connections between the film and the Named Entites are provided as an explanation.

defined set of users, (2) RSS feeds created by portals relevant for our target group, and (3) general news headlines published on Twitter. We continuously update our news database ensuring the news collection is always up to date.

The catalog of available films is imported as a semantically represented data collection. The film data set provides detailed text-based descriptions, including a summary of the plot, the list of related persons (actors, directors, etc.) as well as festivals and awards. In addition to the film catalog we import DBpedia (http://wiki.dbpedia.org/). We use DBpedia Spotlightfor identifying Named Entities in the film descriptions and linking them with their matching DBpedia entity. The results are stored in a semantic store (Blazegraph, https://www.blazegraph.com/).

Processing and Computing Recommendations

The continuously collected news is periodically processed. In the first step we recognize and disambiguate Named Entities in the texts/messages. Subsequently we compute the co-occurrence vectors of each entity using the most recent news documents. The vectors are clustered using the DBScan algorithm [3] in order to obtain clusters representing the topics dominant in the news. The vector processing and clustering is implemented in the Apache Mahout framework (http://mahout.apache.org/). The computed topics are ranked by the number of retweets and references in order to give popular news a bigger weight.

Based on the most popular topics we recommend movies that are closely related. Every entity in a topic is compared to each entity extracted from the film descriptions or the meta-data. The relatedness of films to a topic is computed by the number of entities directly connected to both. In addition, we considered the similarity between linked entities, which is computed based on their attributes by using a vector-space-model (as explained in [2]). In general, the similarity between two entities is the bigger the more they share common subjects (based on DBpedia edges type `skos:subject`). The recommendation result is built from the films containing the Names Entities recognized in the news or entities similar to these entities.

Front End

The recommendations are provided as a ReSTful web service. For visualizing the results we built a web application allowing us testing different parameter sets and collecting feedback from users for the computed recommendations. Fig. 2 shows an example: Based on the Named Entities recognized in a tweet, a movie about a Syrian woman is recommended. One can see how the detected entities contribute to the relevance score.

3. EVALUATION AND SYSTEM DEMO

The implemented web application allows us testing the recommendations and discussing the suggestions.

In the system demonstration we show the recommendations based on live news data. Users can select the sources used for identifying topics. In addition, we demonstrate the effect of different methods used for clustering and computing the similarity between topics and films. The explanations provided for the recommendations help users to understand why a movie is related to the identified topics and why the movie is recommended.

A preliminary evaluation of the recommendation results shows that the implemented recommender approach provides relevant suggestions. The use of topics extracted from politics-related news delivers good recommendations for documentaries and film strongly related to certain regions and crises. News retracted from arthouse-related tweets and RSS feeds are well-situated for detecting artist-related events (e.g. anniversaries and festivals). Entities-based (semantic) similarity measures outperform term-based similarity measures. The weights of the considered semantic edge types must be tuned for the scenario in order to ensure high-quality recommendations.

4. CONCLUSIONS

We presented a novel approach combining topics extraction and linked open data for computing topical recommendations. The preliminary evaluation shows that the recommender delivers highly relevant results. Events detected in the recent news are accepted by users as helpful explanations. As future work, we plan extended user studies. Furthermore, we plan to apply the approach in different domains. The adaption for new domains does not require much effort, since the approach only requires a stream of documents (allowing us identifying Named Entities). When applying the algorithm to new domains the set of considered entities and the similarity measure must be adapted accordingly.

5. REFERENCES

[1] J. Daiber, M. Jakob, C. Hokamp, and P. N. Mendes. Improving efficiency and accuracy in multilingual entity extraction. In *Procs. of the 9th Intl. Conf. on Semantic Systems (I-Semantics)*, 2013.

[2] T. Di Noia, R. Mirizzi, V. C. Ostuni, D. Romito, and M. Zanker. Linked open data to support content-based recommender systems. In *8th Intl. Conf. on Semantic Systems - I-SEMANTICS '12*, NY, USA, 2012. ACM.

[3] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Procs. of KDD-96*, pages 226–231. AAAI, 1996.

Acknowledgments

The work has been partially done in the EEGoF project supported by the German Federal Ministry for Economic Affairs and Energy. The research leading to these results was performed in the CrowdRec project, which has received funding (or partial funding) from the EU 7th Framework Programme FP7/2007-2013 under grant agreement No. 610594.