Data integration for customer preference learning

M. Kopecký[1], M. Vomlelová[2], P. Vojtáš[1]

Faculty of Mathematics and Physics
Charles University
Malostranske namesti 25,
Prague, Czech Republic

[1]{kopecky|vojtas}@ksi.mff.cuni.cz
[2]marta@ktiml.mff.cuni.cz

Abstract. We describe the process and challenges of integration of movie data from Movie Lens, Netflix and RecSys Challenge 2014 with IMDB and DBPedia. Thanks of this integration we can enhance information by semantic data and improve prediction of customer preferences and recommendation. These data were collected in different situation by different methodologies. We want to use these data to be able to extend and further enhance our machine learning approaches developed for individual datasets to other datasets**.**

Keywords: Applications using data extracted from web, computer annotation, data, experiments and metrics

1 Introduction, motivation, recent work.

No human can comprehend any large collection of multi-dimensional data in his/her mind and choose the optimal item according to complex and often difficult to formulate criterion. For this purpose can be helpful recommender systems, that can learn user's preferences from his/her both explicit and implicit actions. The goal of the recommender system is then suggest suitable and often surprising proposals. Different collections of the otherwise similar data can often require different approaches simply due to different semantic data available about items and users in datasets. Because these approaches cannot be directly executed on all datasets, they can be compared only with complications. In this ongoing research report we thus concentrate on synergy effect of annotation and integration of data for user preference learning, and consequently for recommendation. The optimal are such domains where individual items can be identified and where additional data are publicly available. As a basic domain we choose the domain of movies.

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 120-124.

2 Extracting and integrating data from movie domain

In this chapter we first describe data creation, interchanging annotation and data integration. We use *Flix* data i.e. enriched *Netflix* competition data, *RecSys* 2014 challenge data [3] and *RuleML* Challenge data [1].

We started with three available independent datasets: *MovieLens 20M* dataset, *Twitter dataset* and *Flix dataset*

Sizes of all datasets are summarized in Table 1.

Table 1– original datasets

Dataset	Ratings	Rated /all movies	Rating users
MovieLens 20M	20 000 263	26 744 27 278	137 493
Twitter dataset	168 880	13 616 14 542	22 073
Flix dataset	90 217 939	12 031 17 770	479 870

The datasets are quite different. Still they have few things in common. Movies have their title and usually also the year of their production. Ratings are equipped by timestamp that allows us to order ratings from individual users chronologically.

To be able to map movies from different datasets, we wanted to enhance every movie record by the corresponding IMDb[1] identifier **TT** with format ‘ttNNNNNNN’.

We observed that the *Twitter* dataset uses as their internal **MOVIEID** the numeric part of the IMDb identifier. So the movie ”Midnight Cowboy” with **MOVIEID**=64665 corresponds to the IMDb record with ID equal to ’tt0064665’.

To be able to assign IMDb identifiers to movies from other datasets, we had to use the search capabilities of the IMDb database. For both of them we used an HTTP interface for searching movies according to their name. The HTTP response then – among others – contains a table in form:

```
<table><tr>
     <td><a href="/title/ttNNNNNNN/?ref_=fn_ft_tt_1" ><img
src="..."></a></td>
     <td><a href="/title/ttNNNNNNN
/?ref_=fn_ft_tt_1">Title of the movie</a> (YEAR) ...</td>
</tr></table>
```

To be able to maintain both *MovieLens* and *Flix* dataset equally – regardless different formats of movie titles in them – and potentially in other future datasets, we needed to transform each movie title to the proper form expected by the IMDb interface. The basic algorithm can be described in steps:

[1] http://www.imdb.com/

- Convert all letters in movie title to lower case.
- If the movie title contains year of production at its end in brackets remove it.
- If the movie title still contains text in brackets at its end, remove it. This text usually contained original name of movie in original language.
- Move word ”the”, respectively ”a”/”an” from the end of the title to the beginning.
- Translate characters ”_”, ”.”, ”?” and ”,” to spaces
- Translate ”&” and ”&amp;” in titles to word ”and”

For example, the transformation changes title ”Official Story, The (La Historia Oficial) (1985)” from the *MovieLens* dataset to its canonical form ”the official story” which can be identified as movie with the ID=’tt0089276’. Similarly the title ”Seventh Seal, The (Sjunde inseglet, Det) (1957)” from the same dataset is transformed to the form ”the seventh seal” with ID=’tt0050976’.

The successfulness of this approach to map movies from both *MovieLens* and *Flix* datasets is in first line of Table 2.

In optimal case, the table returning from the IMDb search contains exactly one row with the requested record. For this situation the algorithm behaves well and is able to retrieve the correct IMDb identifier. In many other cases the result contained more rows and the correct one or the best possible one had to be identified. For this purpose we enhanced the algorithm by additional steps:

- The correct record should be from the requested year, so the returned table should be searched only for records from this year and other records should be ignored
- The IMDb search provides more levels of tolerance in title matching. Try to use them from the most exact one to the most general. If the matching record from requested year cannot be found using stricter search, the other search level is used.

Currently, we have 13 081 out of all 17 770 *Flix* movies mapped onto the IMDb database. Even all 27 278 out of 27 278 movies from the *MovieLens* set are mapped to the equivalent IMDb records. So the current results provided by the combination of most advanced versions of algorithms are promising.

The diagram in the Figure 1 shows the amount of movies associated to the IMDb record in different intersections after the integration. For each movie registered in the IMDb database we then retrieved XML data from the URL address `http://www.omdbapi.com/?i=ttNNNNNNN&plot=full&r=xml` and then from the XML data we retrieved following movie attributes. Among others *title*, *rating*, *avards*, *year*, *country*, *language*, *genres*, *director* and *actors*.

Another source of semantic data we use is the *DbPedia*. For this purpose we implemented the mapping technique described in [K] and assigned DbPedia[2] identifiers and associated semantic data to *IMDb* movies.

[2] http://wiki.dbpedia.org/

The *DbPedia* identifier of movie is a string, for example ”The_Official_Story” or ”The_Seventh_Seal”. This identifier can then be used to access directly the *DbPedia* graph database or retrieve data in an XML format through the URL address in form `http://dbpedia.org/page/DbPediaIdentifier`.

Table 2 –IMDb search by title name – the successfulness of IMDb title search for original – seven steps – algorithm and the final – enhanced – version.

	MovieLens	**Flix**	**Twitter**
IMDb search by title name	45,4%	70,9%	Not needed
Final enhanced version	100.0%	73.6%	Not needed

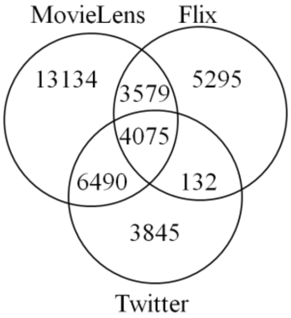

Figure 1 – Integration of movies in datasets based on the IMDb mapping

3 Conclusions, future work

We illustrated our approach to integration of five datasets – three movie datasets and two movie databases containing semantics data.

The future challenge is twofold:

- provide deeper analysis of data mining and use interconnection of datasets and their semantic enhancements for identifying and using possible dataset similarities.
- In future research we would like to continue in approaches in [2].
- extend this approach to other domains

Acknowledgement: Research was supported by Czech project Progres Q48.

References

1. Kuchar J.: Augmenting a Feature Set of Movies Using Linked Open Data, Proceedings of the RuleML 2015 Challenge, Berlin, Germany. Published by CEUR Workshop Proceedings, 2015
2. L. Peska, I. Lasek, A. Eckhardt, J. Dedek, P. Vojtas, D. Fiser: Towards web semantization and user understanding. In EJC 2012, Y. Kiyoki et al Eds. Frontiers in Artificial Intelligence and Applications 251, IOS Press 2013, pp 63-81
3. Twitter data from RecSys 2014 challenge http://2014.recsyschallenge.com/