SemRevRec: A Recommender System based on User Reviews and Linked Data

Iacopo Vagliano
ZBW - Leibniz Information Centre for
Economics, Kiel, Germany
i.vagliano@zbw.eu

Diego Monti
Politecnico di Torino
Turin, Italy
diego.monti@polito.it

Maurizio Morisio
Politecnico di Torino
Turin, Italy
maurizio.morisio@polito.it

ABSTRACT

Traditionally, recommender systems exploit user ratings to infer preferences. However, the growing popularity of social platforms has encouraged users to write textual reviews about liked items. These reviews represent a valuable source of non-trivial information that could improve users' decision processes. In this paper we propose a novel recommendation approach based on the semantic annotation of entities mentioned in user reviews and on the knowledge available in the Web of Data. We compared our recommender system with two baseline algorithms and a state-of-the-art Linked Data based approach. Our system provided more diverse recommendations with respect to the other techniques considered, while obtaining a better accuracy than the Linked Data based method.

1 INTRODUCTION

Currently, most of recommender systems exploit user ratings to infer preferences, although the growing popularity of social and e-commerce websites has encouraged users to write textual reviews. These reviews enable recommender systems to represent the multi-faceted nature of users' opinions and build a fine-grained preference model, which cannot be obtained from overall ratings [2].

In this paper we describe how the information extracted from user reviews, combined with Linked Data, can be exploited in recommendation tasks. On one side, Linked Data can provide a rich representation of the items to recommend since they include interesting features. On the other side, reviews may reveal additional connections among items: for instance, various reviews of *Interstellar* mention Stanley Kubrick, although there is not a direct link between these two resources in DBpedia. We propose a novel recommendation approach based on the semantic annotation of reviews to extract useful information from them. A preliminary offline study suggests that our method provides better prediction and ranking accuracy than another recommender system based on Linked Data, while it increases the diversity of recommendations with respect to all the techniques considered.

2 APPROACH

SemRevRec consists of two main modules: semantic annotation and discovery, and recommendation. The former is responsible for feeding the recommendation module with semantically annotated entities and Linked Data, while the latter provides suggestions to users. The two modules are disconnected: the recommendation module works online, while the other works offline and provides the entities which can be recommended. Every time a new review is submitted, the system can repeat the semantic annotation and discovery steps and possibly identify new entities.

Although our approach is not bounded to a particular domain, in our implementation, we exploited reviews from IMDb[1] because we focused on movies. We chose DBpedia for annotation and discovery because it is one of the main datasets in the Web of Data.

2.1 Semantic Annotation and Discovery

The semantic annotation technique associates a URI to the entities recognized in a given text to add information about their meaning. In our case, the entities identified in the reviews are resources in the Web of Data. Thus, the semantic annotation and discovery module can find other resources that are linked with the annotated entities, in order to enable our system to recommend more items.

In our implementation, we relied on AIDA [3] to annotate reviews with DBpedia resources. We exploited the DBpedia properties `dbo:starring` and `dbo:director` for discovering, through SPARQL queries, additional resources that are connected with the annotated entities. The underling hypothesis is that most of the entities, if not movies, should be actors or directors. However, these properties can be configured according to the domain and the dataset considered. Given the annotated entities, the discoverer retrieves other relevant entities. This allows the system to discover other movies from the same director or actor named in a given review and significantly improve the accuracy of the recommendations. E.g., if Christopher Nolan was annotated in a review of *The Dark Knight, Interstellar* can be found because it is directed by him.

The semantic annotation and discovery module stores both annotated and discovered entities. The URI of each annotated entity is associated with the URI of the reviewed item and with the occurrence of that entity in all the reviews of that item. The same entity may, in fact, appear in reviews regarding different items. Similarly, the URI of each discovered entity is stored together with the URI of the annotated entity through which it was discovered and their Linked Data Semantic Distance (LDSD) [5], a measure inversely proportional to the number of links between two resources.

2.2 Recommendation

The recommendation process consists of two main steps: the generation of the candidate recommendations and their ranking. Given an initial item, SemRevRec retrieves all the entities related to it.

---

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

Firstly, the system selects the annotated entities which were mentioned in the reviews of the initial item. Afterward, it obtains the entities which mention the initial item, i.e. entities whose reviews generated an annotated entity that corresponds to the initial item. For example, if the initial item is *Interstellar* and a review of *2001: A Space Odyssey* mention *Interstellar*, then *2001: A Space Odyssey* is considered as a candidate recommendation.

Subsequently, SemRevRec retrieves the relevant discovered entities. These can be entities discovered through the initial item. For instance, if the initial item is *Interstellar* and *The Dark Knight* was previously discovered because both these movies have been directed by Christopher Nolan, *The Dark Knight* is selected. Similarly, the entities discovered through other entities which were annotated in the reviews of the initial item are relevant. E.g., if *Interstellar* is the initial item, Stanley Kubrick was annotated in one of its reviews, and *2001: A Space Odyssey* was discovered through Stanley Kubrick, then *2001: A Space Odyssey* is a candidate recommendation.

Finally, SemRevRec ranks the candidate recommendations. The ranking function (Equation 1) considers the occurrence $occurrence_i$ of entities in the reviews and the Linked Data Semantic Distance (LDSD) between each discovered entity and the entity through which it was discovered. This avoids assigning the same value to all the entities discovered through the same annotated entity. The item i can be an annotated or a discovered entity. The α coefficient is 1 if the item i is an annotated entity, while it can be configured to a custom value for the discovered entities (by default is 0.5). For the discovered entities, the occurrence of entities through which they were discovered is used, multiplied by α. To obtain a value between 0 and 1, the occurrence is normalized with respect to the maximum occurrence of entities j which belong to the candidate recommendation set CR. The β coefficient is 1 if i is an annotated entity, 0.5 otherwise. The γ coefficient is 0.5 for discovered entities, 0 otherwise. In this way, the function returns a number between 0 and 1, which is equal to the first term for the annotated entities, while, for the discovered entities, it represents the average of the first term and $1 - LDSD(i, i_o)$, where i_o is the entity through which it was discovered.

$$R(i) = \beta \cdot \frac{\alpha \cdot occurrence_i}{max_{j \in CR}(occurrence_j)} + \gamma \cdot (1 - LDSD(i, i_o)) \quad (1)$$

3 EVALUATION

We evaluated SemRevRec with a preliminary offline experiment conducted in the movie domain. Its purpose is to compare our proposal with a state-of-the-art recommender system based on Linked Data and two baseline algorithms. We annotated the reviews available on IMDb for the top-250 movies[2]. We also relied on the MovieLens 1M dataset[3] for obtaining the actual user ratings.

The evaluation was performed with LibRec[4]. We executed a 5-fold cross-validation considering as positive the ratings greater than 3 on a scale from 1 to 5. Using the top-10 recommendations for each user, we computed the measures of precision, recall, nDCG, Entropy Based Novelty (EBN) [1], and diversity [6]. We compared our technique with the Most Popular and the Random Guess baseline algorithms, and with SPrank [4]. We configured SPrank to exploit

[2] http://www.imdb.com/chart/top
[3] http://grouplens.org/datasets/movielens/1m/
[4] https://www.librec.net

Table 1: Results of the experiment

Algorithm	Precis.	Recall	nDCG	EBN	Divers.
SemRevRec	0.0882	0.0459	0.0589	1.5671	0.1838
SPrank	0.0584	0.0327	0.0409	0.8244	0.1551
Popular	0.1325	0.0840	0.0969	2.7439	0.1412
Random	0.0055	0.0028	0.0031	0.3018	0.1679

LambdaMart as ranking method and the properties related to the movie domain (`dct:subject`, `dbo:director`, and `dbo:starring`).

Table 1 lists the results of the experiment. For all the measures but EBN, higher values mean better results, while the lower is EBN, the higher is the novelty. SemRevRec provided a better prediction accuracy and ranking than SPrank, while it improved in novelty with respect to the Most Popular technique. However, SPrank obtained a higher novelty than SemRevRec. The diversity of the algorithms was similar, but our system resulted in the best diversity.

4 CONCLUSIONS AND FUTURE WORK

In this paper we proposed a novel recommendation approach based on the semantic annotation of reviews to extract information as Linked Data. Our method discovers additional resources and generates recommendations by exploiting the annotated entities. A preliminary offline study conducted in the movie domain suggested that our algorithm provides better prediction accuracy and ranking than another method based on Linked Data, while it increases the diversity of recommendations with respect to the other techniques considered. Although we have tested our approach in only one domain, we could apply it to others, provided the reviews. As future work, we plan to evaluate SemRevRec in other domains, such as music and books, and also consider, during ranking, the sentiment and the linking confidence associated with the annotated entities.

ACKNOWLEDGMENTS

REFERENCES

[1] Alejandro Bellogìn, Ivàn Cantador, and Pablo Castells. A Study of Heterogeneity in Recommendations for a Social Music Service. In *Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems (2010) (HetRec '10)*. ACM, 1–8. https://doi.org/10.1145/1869446.1869447

[2] Li Chen, Guanliang Chen, and Feng Wang. 2015. Recommender systems based on user reviews: The state of the art. *User Modeling and User-Adapted Interaction* 25, 2 (2015), 99–154. https://doi.org/10.1007/s11257-015-9155-5

[3] Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust Disambiguation of Named Entities in Text. In *Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, Edinburgh, Scotland*. 782–792.

[4] Tommaso Di Noia, Vito Claudio Ostuni, Paolo Tomeo, and Eugenio Di Sciascio. 2016. SPrank: Semantic Path-Based Ranking for Top-N Recommendations Using Linked Open Data. *ACM Transactions on Intelligent Systems and Technology* 8, 1 (2016), 9:1–9:34. https://doi.org/10.1145/2899005

[5] Alexandre Passant. 2010. dbrec - Music Recommendations Using DBpedia. In *The Semantic Web - ISWC 2010*. Springer Berlin Heidelberg, 209–224.

[6] Mi Zhang and Neil Hurley. 2008. Avoiding Monotony: Improving the Diversity of Recommendation Lists. In *Proceedings of the 2008 ACM Conference on Recommender Systems (RecSys '08)*. ACM, New York, NY, USA, 123–130. https://doi.org/10.1145/1454008.1454030