A method for multidomain entity recommenders

Giorgia Di Tommaso
ditommaso@di.uniroma1.it
Department of Computer Science, Sapienza, University of Rome

ABSTRACT

Recommender systems (RS) are changing the way we buy and discover new products on the web. Many big companies (such as Amazon or Netflix) are now developing RS as part of the services they provide to their users to improve online experience. Despite the presence of many techniques in literature, these systems suffer from some problems, including the cold start problem. The proposed work aims to solve these problems, making an advancement in the state of the art through the implementation of SeREnA an innovative unsupervised RS based on semantics and self-similarity.

KEYWORDS

Semantic recommender systems; users'interests dataset; cold start problem; knowledge base systems, Twitter

1 INTRODUCTION

Generally, RS are classified into three categories based on the filtering method used: *collaborative filtering*[5] generates recommendations aggregating preferences of many users, *content-based filtering* [4] suggests items comparing user's profile and item content, *hybrid approach* combines previous methods. All these methods need to collect preferences of users in order to make accurate recommendations. The need of data on other users and items represents maybe the biggest problem encountered by RS. Insufficient data causes the cold start problem, the incapability to make recommendations of new elements or to new users when there is not enough knowledge on previous usage of the RS. Other weaknesses [3] affecting these methods are sparsity, volatility and many others. Recently, a new generation named *knowledge-based RS* has emerged to alleviate these deficiencies. In this context our work SeRenA aims at developing a RS based on semantics that is able to exploit available information from a single user, such as his/her online activities on social networks, to recommend items in any domain (eg. sport music, food and others). SeRenA uses Wikipedia as knowledge base to categorize data about users'interests.

2 METHODOLOGIES

Our approach is composed by the three following steps:

(1) *Extraction of users' interests and mapping onto related instances of a Knowledge Base:* To build knowledge about a user the first step is the extraction of interests from social media. We use Twitter because of its availability under specific restrictions. We extract multi-domain preferences from a user's messages and from his/her friendship list, identifying those followees who represent an interest rather than a peer friendship relationship (e.g. celebrities, places..). Then we map these preferences onto Wikipages (when possible).

(2) *Inferring the Semantic Fingerprint:* Given a set of Wikipedia pages, we extract a set of connected *categories* that summarize at best (with the appropriate level of generality/specificity) a user's interests. We denoted this set as Semantic Fingerprint $SF(u)$. We use text mining techniques to extract an initial set of seed categories then we apply graph pruning and word embedding algorithms to generate $SF(u)$. This is the most challenging part of the proposed project, since large knowledge bases have many undesired properties, such as cycles, multiple inheritance, and noise.

(3) *Recommending entities:* Finally, we use $SF(u)$ to recommend u with new items. We propose a measure called *propensity* $P(e, u)$ of a user u for a new entity e. The propensity can be estimated as a function of the weight of categories in the fingerprint, and their (shortest) distance from e.

3 RESULTS AND CONCLUSION

To evaluate SeRenA, we define offline experiments and test the system in several critical settings: recommending cold start users and items. We compute a list of metrics such as Precision, Accuracy, Reciprocal Rank under different configurations (varying number of users and preferences). Using a specially crafted dataset [2], the applicability of SeRenA in different domains and sources was studied. Our results were compared with methods in the state of the art both semantic-based and not. For the sake of space, we cannot show results for each experiment, but Figure 1 shows that in a cold start environment the proposed method using semantic and self similarity is more efficient in term of precision than the closest method based on collaborative filtering and social data [1].

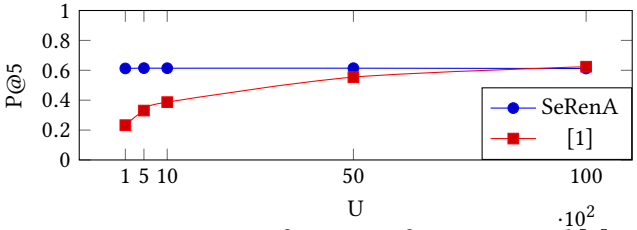

Figure 1: Comparison of Precision for SeRenA and [1]

ACKNOWLEDGMENTS

This work is conducted under the supervision of Prof. Paola Velardi and Dr. Giovanni Stilo, in collaboration with Dr. Stefano Faralli.

REFERENCES

[1] Allison JB Chaney, David M Blei, and Tina Eliassi-Rad. 2015. A probabilistic model for using social networks in personalized item recommendation. In *Recsys'15*. ACM, 43–50.
[2] Giorgia Di Tommaso, Stefano Faralli, Giovanni Stilo, and Paola Velardi. 2018. WikiMID: a very large Multi-domain Interests Dataset of Twitter users with mappings to Wikipedia. In *ISWC*. Springer, 36–52.
[3] Shah Khusro, Zafar Ali, and Irfan Ullah. 2016. Recommender systems: Issues, challenges, and research opportunities. In *ICISA 2016*. Springer, 1179–1189.
[4] Michael J Pazzani and Daniel Billsus. 2007. Content-based recommendation systems. In *The adaptive web*. Springer, 325–341.
[5] J Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen. 2007. Collaborative filtering recommender systems. In *The adaptive web*. Springer, 291–324.