Item Contents Good, User Tags Better: Empirical Evaluation of a Food Recommender System

David Massimo
Free University of Bozen-Bolzano
Bozen-Bolzano, Italy
david.massimo@stud-inf.unibz.it

Mehdi Elahi
Free University of Bozen-Bolzano
Bozen-Bolzano, Italy
meelahi@unibz.it

Mouzhi Ge
Faculty of Informatics, Masaryk University
Brno, Czech Republic
mouzhi.ge@muni.cz

Francesco Ricci
Free University of Bozen-Bolzano
Bozen-Bolzano, Italy
fricci@unibz.it

ABSTRACT

Traditional food recommender systems exploit items' ratings and descriptions in order to generate relevant recommendations for the users. While this data is important, it might not entirely capture the true users' preferences. In this paper, we analyse the performance of a food recommender that allows users to enter their preferences in the form of both ratings and tags, which are then used by a Matrix Factorization (MF) rating prediction model. The performed offline and online experiments have clarified the importance of user tags in comparison to content features. While item content contributes more to the quality of the prediction accuracy, user tags yields better ranking quality. Even more importantly, a live user study has revealed that a system variant, which leverages user tags in the prediction model and in the interface, achieves a significantly better user evaluation in terms of perceived effectiveness, choice satisfaction and choice difficulty.

1 INTRODUCTION

A variety of food recommender systems have been recently developed and evaluated [2, 6, 7]. However, they mostly implement the traditional content-based approach, which leverages the item descriptions (e.g., food categories and ingredients) that a user preferred in the past, in order to generate relevant and new recommendations for her. While content-based solutions could obtain a satisfactory level of recommendation quality, we conjecture that they might not fully model the users' specific preferences and tastes.

More novel recommendation algorithms have integrated tag data into matrix factorization, as additional features of the items. For instance, in [1] the authors proposed a modified version of the SVD++ matrix-factorization model that makes use of tagging information and achieves a substantial improvement of the recommender system performance. Along with this novel line of research, in [4] we have presented a food RS that leverages tags in the predictive model and in the human/computer interaction.

In this paper, we present the result of the offline and online experiments conducted on the system described in [4] in order to better understand the role that tags in preference elicitation and recommendation. The conducted offline experiments show that while item content features contribute more to the quality of the rating prediction accuracy of the system, tags are more useful for obtaining higher quality ranking. Moreover, the online evaluation shows that a system variant that uses tags in the prediction model and in the human/computer interaction achieves a better performance in terms of user perceived recommendation quality, choice satisfaction and choice difficulty.

2 SYSTEM PROTOTYPE

The proposed recommendation model is implemented in an Android-based app. The humam-computer interaction has been designed to support a user who would like to cook at home.

The user-system interaction process begins with the initial sign-up, where the user enters her age and gender. Then, she goes through the preferences elicitation steps, where she is first requested to enter her general preferences by specifying the recipes she eats or cooks at home. Afterward, in a second step, a list of selected recipes are presented to the user for rating. The user is also requested to "explain" the core motivation for an assigned rating, by optionally adding to a rated recipe some tags. In the supported interaction, the system gives two tagging possibilities to the user: (a) marking any ingredient as a tag, (b) adding other free tags that she deems as relevant to the recipe. At recommendation time the user is asked to provide session-based preferences, i.e., the ingredient (e.g., tuna fish) that she is willing to cook.

The full set of the previously entered ratings and tags are used to train the model described in [4] and the generated recommendations are post-filtered to suit the session-based preferences of the user. The system presents, one by one, recipe recommendations to the user and requests her to make a choice. Once the choice is made, the system presents the cooking instructions of the chosen recipe.

We have developed and evaluated various recommendation algorithms [4]. All of them implement rating-based MF (Matrix Factorization) models, extended to use different sources of data:

- **MF Content** extends MF by using item content features;
- **MF UTag** extends MF by using user tags;

UMAP'17, July 9–12, 2017, Bratislava, Slovakia
© 2017 Copyright held by the owner/author(s). ACM ISBN 978-1-4503-4635-1/17/07.
DOI: http://dx.doi.org/10.1145/3079628.3079640

- **MF UTag Content** extends MF by using both item content features and user tags.

3 EVALUATIONS AND RESULTS

The used recipe dataset has been obtained from *Total Wellbeing Diet* [3] and from authoritative online cook communities: *Food52* and *All recipes*. It contains 234 recipe items, each one described by features, such as, cooking instructions, ingredient list, dish categories they belong to. These recipes were rated and tagged by users who tried the system in several occasions.

We have collected 392 ratings and 394 tags (out of 120 unique tags) provided by 43 users, who participated in our experiment. The rating sparsity of the dataset is 96%. On average, the users rated 9.11 recipes. The number of item content features per rating (i.e., 9.14) is almost an order of magnitude larger than the number of user tags per rating (i.e., approximately 1).

In an offline experiment, for every user, a train set is created by considering 80% of her ratings as well as the entire set of the ratings of the other users. The remaining 20% of the ratings of that particular user are included in the test set. This process is iterated over the entire collection of users and the performance metrics, which are computed for every user, are then averaged. The results show that:

- **MF content** has MAE=0.814 (Mean Absolute Error), which is the best rating accuracy prediction and it is the second best in ranking accuracy, MRR=0.469 (Mean Reciprocal Rank);
- **MF UTag Content** is the second best in terms of rating accuracy prediction (MAE=0.866), but it is the best in terms of ranking accuracy (MRR=0.501);
- **MF UTag** has MAE=0.979, i.e., the worst rating accuracy prediction, but it is still better than plain MF (MAE=0.972), and has the worst ranking accuracy prediction (MRR=0.463), but again, still better than MF (MRR=0.416).

In a live user experiment, 31 participants were randomly assigned to two groups, each evaluating different version of the full Android mobile app recommender system. A first system version, which we call T, allows users to express preferences in the form of ratings and tags, while a second version, named as R, allows preferences' expression only in terms of ratings. These two versions generate recommendations by using the first and the second best performing models (according to the offline results). Indeed, the model used by T is MF UTag Content, which uses ratings, user tags and item content. The model used by system R, on the other hand, is MF Content, which uses user ratings and item content. The number of participants in the first group (T) was 12, whereas in the second group (R) was 19.

Each subject went through the whole user-system interaction process and then answered to a questionnaire, designed and validated by Knijnenburg [5], to measure: (1) perceived system effectiveness, (2) choice satisfaction and (3) choice difficulty. User's responses for the questions were averaged and the differences in performance where compared for statistical significance by using the Mann-Whitney test.

Significant differences between the two systems (p-value<0.05) have been observed for the following statements, among all those included in the questionnaire:

- *I make better choices with system* (perceived effectiveness);
- *I like the items I've chosen* (choice satisfaction);
- *I changed my mind several times before making a decision* (choice difficulty).

The first two statements are positively formulated (the higher the better), while the third statement is negatively formulated (the lower the better). The average responses of the participants to these statements, for system T are 3.7, 4.4, and 2.0, which are significantly better than the average responses of the system R users, with values 3.1, 3.7, and 2.4. Hence, the users believe that the system T shows recommendations that are more useful and more helpful in supporting them to make better choices, in comparison to the recommendations shown by system R. Moreover, the users liked more the recommended items presented by the system T than those listed by R. Finally, users evaluated the system T superior to the system R in enabling them to a make quick decisions without changing their minds several times.

4 CONCLUSION AND FUTURE WORKS

In this paper, we have investigated the role of tags in improving the quality of a food recommender system, in terms of rating prediction accuracy, ranking quality, perceived system effectiveness, and choice satisfaction and difficulty. The performed experiments show the advantages and the disadvantages of using tags in food recommendation.

In a future work, we would like to replicate the designed experimental study on a larger dataset, with a lower level of sparsity and more tag assignments. We believe that more tag assignments can allow to better understand the effects of tags. Further studies should also be carried out in order to study the temporal evolution of tagging patterns and user satisfaction over time.

REFERENCES

[1] Manuel Enrich, Matthias Braunhofer, and Francesco Ricci. 2013. Cold-Start Management with Cross-Domain Collaborative Filtering and Tags. In *Proceedings of the 13th International Conference on E-Commerce and Web Technologies*. Springer, 101–112.

[2] Jill Freyne and Shlomo Berkovsky. 2010. Intelligent food planning: personalized recipe recommendation. In *Proceedings of the 15th International Conference on Intelligent User Interfaces*. ACM, 321–324.

[3] Jill Freyne and Shlomo Berkovsky. 2013. Evaluating recommender systems for supportive technologies. In *User Modeling and Adaptation for Daily Routines*. Springer, 195–217.

[4] Mouzhi Ge, Mehdi Elahi, Ignacio Fernaández-Tobías, Francesco Ricci, and David Massimo. 2015. Using Tags and Latent Factors in a Food Recommender System. In *Proceedings of the 5th International Conference on Digital Health 2015 (DH '15)*. ACM, New York, NY, USA, 105–112. http://doi.acm.org/10.1145/2750511.2750528

[5] Bart P Knijnenburg, Martijn C Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. 2012. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction* 22, 4-5 (2012), 441–504.

[6] Tomasz Kusmierczyk, Christoph Trattner, and Kjetil Nørvåg. 2015. Temporality in online food recipe consumption and production. In *Proceedings of the 24th International Conference on World Wide Web*. ACM, 55–56.

[7] Michele Trevisiol, Luca Chiarandini, and Ricardo Baeza-Yates. 2014. Buon appetito: recommending personalized menus. In *Proceedings of the 25th ACM conference on Hypertext and social media*. ACM, 327–329.