Increasing Top-20 search results diversity through recommendation post-processing

Matevž Kunaver[1], Štefan Dobravec[1], Tomaž Požrl[1], and Andrej Košir[1]

[1]University of Ljubljana, Faculty of Electrical Engineering, Tržaška 25, Ljubljana, Slovenia
matevz.kunaver@fe.uni-lj.si,stefan.dobravec@fe.uni-lj.si,,tomaz.pozrl@fe.uni-lj.siandrej.kosir@fe.uni-lj.si

Abstract. This paper presents three different methods for diversifying search results, that were developed as part of our user modelling research. All three methods focus on post-processing search results provided by the baseline recommender systems and increase the diversity (measured with ILD@20) at the cost of final precision (measured with F@20). The authors feel that these methods have potential yet require further development and testing.

Keywords: Recommender systems, Diversity, ILD, F-measure, User modelling

1 Introduction

The focus of recommender systems (RSs) is moving from generating recommendations (providing personalized data retrieval and search results) without any additional situational data about the user to generating recommendations that also consider the user's context [1][3] and personality in order to improve the recommendation results[7]. All these improvements serve to present the user with a selection of results that will be the most appropriate for the situation in which the user desires to review the selected result. Recommendation results can be further improved by paying attention to the diversity [4] [8] [5] [11] of results presented to the user.

1.1 Motivation and Goal

The purpose of our study is to determine whether we can increase the diversity of results generated and presented to the user by a baseline RS by introducing three methods that post-process these results. Each of these methods uses a different diversification approach yet all three aim to maintain a high level of user satisfaction (measured by evaluating the accuracy of the modified RS). While 'search results' cover a wide array of possible items, we focused our research on movie search results as we had two different working RSs developed as part of our previous research in movies domain [9][10] and could therefore immediately focus on diversification method development.

2 Materials and Methods

In this section we describe the dataset, the baseline RSs used to generate recommendations, the developed diversification methods and the evaluation methods.

2.1 Dataset

For the purposes of our research we used the *Context Movie Dataset* (LDOS-CoMoDa), that we have acquired in our previous work. The dataset was collected using an on-line application for rating movies (www.ldos.si/recommender.html) that enabled the users to track the movies they have watched and to obtain recommendations from several RS algorithms. In addition, the application features a questionnaire whose purpose is the collection of the contextual data describing the situation during the item consumption.

The dataset currently consists of 4237 ratings given by 184 users to 1782 items. Each rating is also annotated with associated contextual variables. Each user is described with basic demographic data (age, sex, location) provided on a voluntary basis. Each item is described with several attributes: genre, director, actor, language, country, budget and release year.

The on-line application is still available and in use. Additional information about LDOS-CoMoDa can be found in [2] and [3].

2.2 Recommender System

For this paper we implemented our diversification methods on two different RSs: a hybrid RS and a content-based RS.

Hybrid RS[9]: The hybrid RS used for this experiment was developed as part of our previous research [9]. It is a collaborative RS that selects nearest neighbours based on genre preferences instead of their ratings. Each preference indicates the user's interest for one specific genre (25 in total). By using these preferences we are able to select nearest neighbours who perfectly match the active user in preferences without having a single overlapping item (i.e. item rated by both users). This increases the recommendation pool and the overall quality of the RS.

The hybrid RS generates recommendations for each user by performing the following steps: (i) Calculate genre preferences for the user based on his/her previous ratings, (ii) Find 20 users whose preferences are the most similar to the active user, (iii) Create a pool of potential recommendations from all of the items rated by these users, (iv) Calculate the predicted rating for each item using the Bayesian estimator, (v) Present the user with the top 20 items.

Content-based RS[10]: The content-based RS used in this paper developed as part of our previous research [10] as well and is based on a rule-based approach that considers all attributes available in the dataset. We defined a special similarity function that enables us to detect attribute values in the description of the item that user has a very high preference towards. If we detect

such attribute values, we assign a high similarity value between the attribute value and the model of the user.

The content-based RS performs the following steps: (i) Generate content-based user model from items the user has already watched and rated, (ii) For each item not yet rated calculate attribute similarity values for attributes in item metadata using content-based user model, (iii) First, calculate similarity for each attribute value and then combine these similarities to a similarity of the attribute, (iv) Classify a vector of similarities of attributes into one of the rating values using 'M5Rules' decision rule classification method.

2.3 Diversification method

We aimed to develop methods that could be implemented in existing RSs without requiring a direct change in the way those RSs work. We therefore focused on diversifying the top 20 lists generated by those RSs. In our case the diversification process is following next steps (as shown in figure 1): for every user's list of recommended items (i) prepare ordered (descending) list of recommendations and split it into top 20 recommendations list and the remainder of the set, (ii) find exchange candidates in such manner that the diversity of top 20 items increases without significant harm to the accuracy of the system, and (iii) exchange the items to yield diversified list of recommended items. As indicated in figure 1 the second and the third step can be performed iteratively.

In our experiment we developed and tested three variations of the diversification process that differ mainly in the way how the exchange candidates are picked.

The first method swaps up to three items in a single step (no iteration). It starts by assessing the worst items in the top 20 list. It calculates the ILD value of the list while excluding one item (exchange candidate) of the list at a time. Effectively this means calculating ILD@19. Higher values of ILD@19 indicate better exchange candidates. Next, it searches for the best replacement candidates from the first 20 items of the remainder of the set. The method calculates the ILD@20 after exchanging every combination of up to three items. Final result is the top 20 list with best ILD@20 score after the exchange.

The second method uses the same approach as the first one to determine exchange candidates in the top 20 list. The best item, which yields highest ILD@19 is then replaced with an item from the first 20 of the remainder of the set. The final exchange is done using the replacement candidates that gives best ILD@20 score. In this case we shuffle only a single item at a time, but repeat the process K-times. It can be expected that increasing value of K would favour list diversity in trade-of to lowering list accuracy.

The third method, just like the second one, replaces single item at a time. The difference is, it considers a joint score in form of $a * avgPR + b * nILD$ instead of a pure ILD value. In this formulation $avgPR$ stands for the average prediction rating of the list and $nILD$ for the normalized ILD value of the same list. Parameters a and b allow balancing the top 20 list from more accurate / less

diverse towards less accurate / more diverse. The shuffling procedure is repeated until best top 20 list (in term of joint score) is achieved.

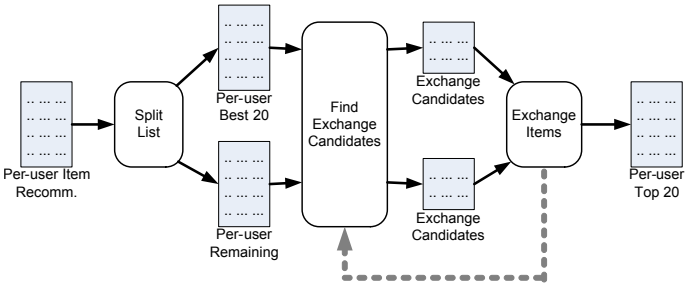


Fig. 1: Diversification process

2.4 Evaluation methods

In order to evaluate our methods and compare them to the control (non-diversified) RS we had to consider accuracy as well as diversity of each generated top list.

We evaluated **accuracy** using the F-measure at top ranking position 20 (F-measure@20) [6] as it is one of the most often used measures of accuracy in recommender systems. In order to evaluate the **diversity** of our recommendations we used the intra-list diversity [11] (ILD@20) calculating the diversity value of each top list based on the following metadata descriptions of each item: genre, director, actor, language and country.

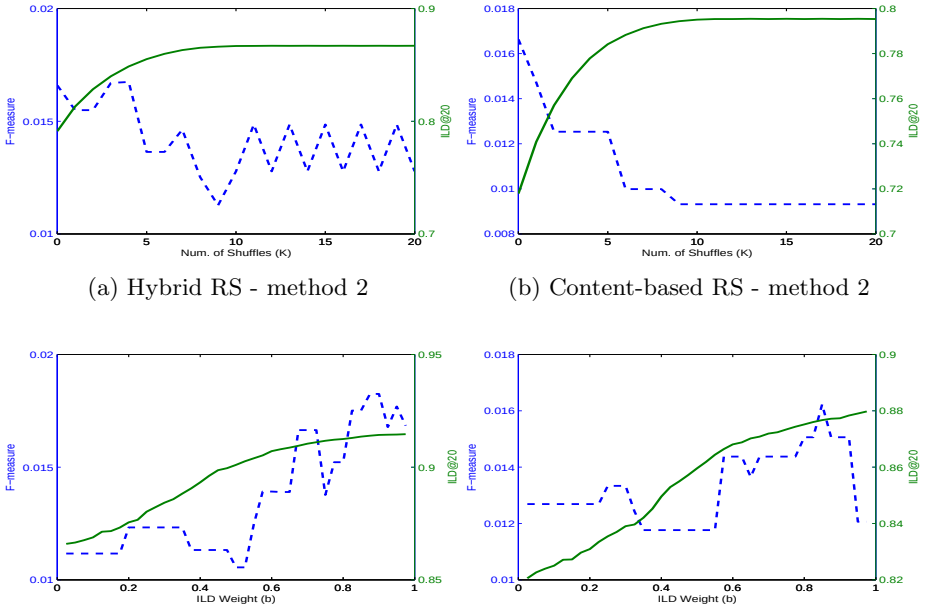
3 Results

Table 1 shows the results of our evaluation (F@20 and ILD@20) of both baseline recommender systems and for all three developed diversification methods.

Table 1: Evaluation results

	Hybrid RS		Content-based RS	
method	F@20	ILD@20	F@20	ILD@20
non-diversified	0.011	0.772	0.020	0.717
diversified - method 1	0.007	0.818	0.0122	0.764
diversified - method 2	0.015	0.867	0.0125	0.784
diversified - method 3	0.018	0.915	0.0151	0.878

As methods 2 and 3 featured additional parameters (number of iterations / joint score settings) we also present their results in figure 2, where we show how different parameter settings impact the systems accuracy / diversity.



(a) Hybrid RS - method 2

(b) Content-based RS - method 2

(c) Hybrid RS - method 3

(d) Content-based RS - method 3

Fig. 2: Evaluation results - method 2 and 3

4 Conclusion and Further Work

The results presented in this paper show promise as all three diversification methods increased the overall top 20 list diversity by at least 6% with the best increase being by method 3 which increased the diversity of content-based recommendations by 22%. The main difference between all three methods is that the first method is a non-iterative one and therefore requires a single run to diversify all top 20 lists while methods 2 and 3 require several iterations to provide the best results in addition to requiring and extra training run to determine the best parameter values.

The real surprise however came when we measured the impact on accuracy for each method. While we saw a decrease in accuracy in content-based RS (from 25% to 40% as expected) we actually found that diversifying our hybrid RS with method 2 or 3 increases the overall accuracy by as much as 60%. We think that this might be due to the small number of ratings per user in our dataset (meaning that shuffling the top items managed to hit a few additional items in the test set, thus increasing the R@20 and P@20 values) and that using the same method on a different dataset might yield different results. However, we also believe that we should use additional accuracy evaluation methods in our future experiments and see if they support the findings from this paper.

We have nevertheless started a study in post-processing diversification that shows promise and we plan to further expand our understanding by addressing these key issues:

- Determine whether the number of replaced items from the top list can be fixed or must be calculated iteratively for each user each time the RS generates recommendations.
- The number of replacement candidates to be considered.
- Perform a series of statistical tests in order to determine whether our results are really significantly different from those of a non-diversified RS.
- Determine the optimal values of parameters a and b for the third method.
- Perform an A/B test to determine how the lower accuracy impacts the actual user satisfaction.
- Perform a study of method efficiency to determine which of the three methods performs best in which circumstances - when can we afford the extra iterations required by methods 2 and 3 and when we cannot.

References

1. Dey A. and Abowd G. Towards a better understanding of context and context-awareness. pages 304–307, 199.
2. Košir A., Odic A., Kunaver M., Tkalcic M., and Tasic J. F. Database for contextual personalization. *Elektrotehniški vestnik*, 78(5):270–274, 2011.
3. Odic A., Tkalcic M., and Košir A. Managing irrelevant contextual categories in a movie recommender system. *Human Decision Making in Recommender Systems (Decisions@ RecSys 13)*, page 29, 2013.
4. Smyth B. and McClave P. Similarity vs. diversity. In *Case-Based Reasoning Research and Development*, pages 347–361. Springer, 2001.
5. Jannach D., Lerche L., Gedikli F., and Bonnin G. What recommenders recommend–an analysis of accuracy, popularity, and sales diversity effects. In *User Modeling, Adaptation, and Personalization*, pages 25–37. Springer, 2013.
6. Ricci F., Rokach L., Shapira B., and Kantor P.B. *Recommender Systems Handbook*. Springer, 2011.
7. Adomavicius G. and Tuzhilin A. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734–749, 2005.
8. Adomavicius G. and Kwon Y. Improving aggregate recommendation diversity using ranking-based techniques. *Knowledge and Data Engineering, IEEE Transactions on*, 24(5):896–911, 2012.
9. Kunaver M., Košir A., and Tasič J. F. *Hybrid recommender for multimedia item recommendation*. Lambert Academic Publishing, 2011.
10. T Požrl, M Kunaver, M Pogačnik, A Košir, and JF Tasič. Improving human-computer interaction in personalized tv recommender. *International Journal of Science and Technology, Transactions of Electrical Engineering*, 36(E1):19–36, 2012.
11. Cai-Nicolas Ziegler, Sean M McNee, Joseph A Konstan, and Georg Lausen. Improving recommendation lists through topic diversification. In *Proceedings of the 14th international conference on World Wide Web*, pages 22–32. ACM, 2005.