Expert Systems With Applications 81 (2017) 321–331

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Efficient optimization of multiple recommendation quality factors according to individual user tendencies

Michael Jugovac*, Dietmar Jannach, Lukas Lerche

TU Dortmund, Department of Computer Science, Otto-Hahn-Str. 12, 44227, Dortmund, Germany

ARTICLE INFO

Article history:
Received 11 July 2016
Revised 23 March 2017
Accepted 24 March 2017
Available online 29 March 2017

Keywords:
Recommender systems
Quality factors
User-specific optimization
Trade-offs

ABSTRACT

Recommender systems are among the most visible applications of intelligent systems technology in practice and are used to help users find items of interest, for example on e-commerce sites, in a personalized way. While past research has focused mainly on accurately predicting the relevance of items that are unknown to the user, other quality criteria for recommendations have been investigated in recent years, including diversity, novelty, or serendipity. Considering these additional factors, however, often leads to the following two challenges. First, in many application domains, trade-offs like "diversity vs. accuracy" have to be balanced. Second, it is not always clear *how much* diversity or novelty is desirable in practice.

In this work, we propose a novel parameterizable optimization scheme that re-ranks accuracy-optimized recommendation lists in order to cope with these challenges. Our method is both capable of considering multiple optimization goals at the same time and designed to consider *individual user tendencies* regarding the different quality factors, like diversity. In contrast to previous work, the method is not restricted to a specific underlying item ranking algorithm and its generic design allows the algorithm to be parameterized according to the requirements of the application domain. Experimental evaluations with different datasets show that balancing the quality factors with our method can be done with a marginal or no loss in ranking accuracy. Given that our method can be applied in various domains and within the narrow time constraints of online recommendation, our work opens new opportunities to design novel finer-grained personalization approaches in practical applications.

1. Introduction

Automated and personalized recommendations have become an integral part of our online user experience. Nowadays, recommender systems (RS) are used to help users find relevant items in a variety of ways, e.g., by recommending items to purchase on e-commerce sites, music to listen to on streaming platforms, or other people to connect with on social networks. RS are among the most successful applications of intelligent systems in practical environments. As a consequence, RS are an active field of research and remarkable advances were made in recent years in terms of increasing the prediction and ranking accuracy of recommendation algorithms using, for example, novel methods for matrix factorization, ensemble learning, or learning-to-rank.

Predicting the relevance of an item for a user as accurately as possible is an important quality criterion of a recommender system. It is, however, not the only one (Jannach, Resnick, Tuzhilin, & Zanker, 2016; McNee, Riedl, & Konstan, 2006). Even when the recommended items match the user's taste very well, their value for this user might be limited, e.g., when the recommendations are too obvious and contain only the most popular items. In addition, the resulting recommendation lists can also be too monotonous and their limited diversity might prevent the user from discovering further relevant items or item categories. Several proposals have been made over the last years to deal with these problems, e.g., by increasing the diversity of the recommendations, by counteracting the popularity bias of algorithms and pushing novel, long-tail items, or by making serendipitous suggestions, see, e.g., Adomavicius and Kwon (2012); Iaquinta et al. (2008); Said, Fields, Jain, and Albayrak (2013); Zhang and Hurley (2008); Zhang, Séaghdha, Quercia, and Jambor (2012); Ziegler, McNee, Konstan, and Lausen (2005).

However, two main issues arise when additional quality factors are to be taken into account. First, in many application domains trade-offs have to be balanced. For example, when trying to help users discover new artists on a music streaming platform with a recommender system, one can recommend tracks of artists that

* Corresponding author.
E-mail addresses: michael.jugovac@tu-dortmund.de (M. Jugovac), dietmar.jannach@tu-dortmund.de (D. Jannach), lukas.lerche@tu-dortmund.de (L. Lerche).

http://dx.doi.org/10.1016/j.eswa.2017.03.055
0957-4174/

the user has not yet listened to. However, recommending more novel or lesser-known items can be risky and even detrimental to the perceived quality of the service (Chau, Ho, Ho, & Yao, 2013; Ekstrand, Harper, Willemsen, & Konstan, 2014). Recommending generally popular items, on the other hand, might be algorithmically easier (Steck, 2011) and less risky, but can also be of little value to users who are interested in discovering something new. One algorithmic problem is, therefore, to find a balance between different quality factors, e.g., to increase the diversity of the recommendations without including too many items that are of limited relevance to the user as done, e.g., by Adomavicius and Kwon (2012).

Second, most of the existing works assume that the same global level for each quality factor, such as diversity, is appropriate for the whole user base. For example, the approach of Adomavicius and Kwon (2012) is designed to always *maximize* diversity for all users. In reality, however, the "right value" for each quality factor can depend on many aspects including *the domain* (homogeneity might be more desirable than diversity, e.g., in the music domain), *the individual user* (some users simply prefer a small subset of item categories), or *business goals*, e.g., the promotion of long-tail items.

In this paper, we present a novel algorithmic approach to address these two issues. At its core, our approach consists of a generic optimization scheme designed to balance accuracy with one or more quality factors. In contrast to previous approaches, e.g., by Oh, Park, Yu, Song, and Park (2011), Vargas and Castells (2011), Adomavicius and Kwon (2012), or Kapoor, Kumar, Terveen, Konstan, and Schrater (2015), our method is capable of dealing with multiple optimization goals in parallel and is not limited to simple characteristics like the popularity of the individual items. Furthermore, our approach does not require the use of a specific algorithm for relevance-based item ranking, but is based on re-ranking the topmost items of an accuracy-optimized recommendation list that can be generated by any rating prediction or item ranking algorithm.

To deal with the problem of determining the right level for factors like popularity, we look at the *tendencies of the individual users* based on their past behavior and emulate these tendencies in the re-ranking process. If, for example, a user always had a strong preference for blockbuster movies from a wide range of genres, the re-ranked recommendations will reflect this by putting an emphasis on popular movies and high genre diversity. By using such an individualization, we can avoid the need for the specification of a "global" level of a certain quality feature and, at the same time, better match the user's preferences regarding quality factors that concern the set of recommendations as a whole.

Overall, the general nature and domain independence of our method allow it to be used in various application domains of recommender systems in which trade-offs have to be balanced. Technically, due to the greedy nature of the optimization scheme, it can be applied under the narrow time constraints of online recommendation and used as a post-processing step that further optimizes the outputs of today's high-performance machine learning algorithms. As a result, our method can help to build next-generation recommender systems that apply personalization not only when determining the relevance of individual items, but also on an aggregate level of an entire recommendation list.

The paper is organized as follows. In Section 2, we first discuss technical preliminaries regarding possible approaches of modeling past user tendencies and then present the proposed generic adaptation procedure. In Sections 3 and 4, we report detailed results of a comprehensive set of empirical evaluations in which we test our approach with different data sets, compare it with previous approaches, and investigate system-wide effects of tuning the recommendations to individual user tendencies. Section 5 compares our work on a theoretical level with past approaches that have similar goals.

2. Technical approach

2.1. Preliminaries

Generally, we assume that we are given a set of users with known past preferences, e.g., expressed through explicit item ratings, and a finite set of recommendable items. Our overall goal is to shape recommendations that reflect the user's preference tendencies in one or more quality dimensions while keeping accuracy high. With accuracy we refer to an algorithm's capability of estimating the relevance of an item for a user, which is typically quantified in the literature through computational metrics like precision, recall, or the RMSE.

To be able to quantify how well a recommendation list matches a user's past tendencies with respect to different quality characteristics, we introduce two functions: $\mathcal{P}(S_u)$ represents the user's preference tendencies, whose calculation is based on a sample set S_u, which is a subset of representative items of the user u. S_u can, for example, contain the set of top-rated items of the user, the set of recently listened music tracks, or any other set of explicit or implicit ratings considered to be relevant in the user's current context. $\mathcal{R}(T_u)$, in turn, represents the characteristics of the recommendations, whose values are based on T_u, the ranked top-n list of items recommended to u. This list can, as mentioned above, be generated by any existing recommendation algorithm.

2.1.1. Quantifying the tendencies

Different approaches are possible to numerically quantify a user's past tendencies. In the following, we discuss three alternatives. However, our optimization method is not limited to these approaches and alternative and domain-specific measures can be used.

Mean and Standard Deviation. A straightforward way of quantifying the tendencies is to look at average values and standard deviations of a quality factor. For example, if a user likes blockbusters, the user's sample set will exhibit a high average popularity. In this case, we can try to reflect this tendency in the recommendation list by including items so that the average item popularity of the list is also high. To avoid that a re-ranking algorithm includes too many "extreme" values in the top-n list when trying to match the user's average value, we propose to use scores that combine the mean and standard deviation in a meaningful form, e.g., by using a weighted combination of their absolute values.

Aggregate Measures. Measures, such as the Intra-List-Similarity (ILS), also result in one single aggregated numerical score but cannot be obtained by averaging individual item features. Instead, the ILS is calculated as the mean of the pairwise item similarities. Another example of a measure of this type is the "coherence" of a music playlist, which can, e.g., be determined by looking at the tempo difference of two consecutive tracks. In our evaluations, we use the ILS score (Ziegler et al., 2005) to compare user tendencies and recommendation lists, i.e.,

$$\mathcal{P}_{sim}(S_u) = \sum_{i \in S_u} \sum_{j \in S_u} \frac{sim(i, j)}{|S_u|^2}$$

where $sim(i, j)$ is computed using the cosine similarity of the *Term Frequency-Inverse Document Frequency* (TF-IDF) representations of the items' content descriptions.

Earth Mover's Distance. The re-ranking approach proposed by Oh et al. (2011) shares goals with our work. The authors propose to use the *Earth Mover's Distance* (EMD) to compare two distributions, because the EMD has certain advantages over the Kullback–Leibler divergence in that context. Informally speaking, the EMD