Exploiting Non-content Preference Attributes through Hybrid Recommendation Method[*]

Fernando Mourão[1], Leonardo Rocha[2], Joseph Konstan[3], Wagner Meira Jr.[1]

[1]Universidade Federal de
Minas Gerais
Computer Science
Belo Horizonte, MG, Brazil
{fhmourao,meira}@dcc.ufmg.br

[2]Universidade Federal de São
João Del Rey
Computer Science
São João Del Rey, MG, Brazil
lcrocha@ufsj.edu.br

[3]University of Minnesota
Computer Science and
Engineering
Minneapolis, MN, USA
konstan@cs.umn.edu

ABSTRACT

This paper explores a method for incorporating into a recommender system explicit representations of user's preferences over non-content attributes such as popularity, recency, and similarity of recommended items. We show how such attributes can be modeled as a preference vector that can be used in a vector-space content-based recommender, and how that content-based recommender can be integrated with various collaborative filtering techniques through reweighting of Top-M recommendations. We evaluate this approach on several recommender systems datasets and collaborative filtering methods, and find that incorporating the three preference attributes can lead to a substantial increase in Top-50 precision while also enhancing diversity and novelty.

Categories and Subject Descriptors

H.4.m [**Information System Applications**]: Miscellaneous; H.m [**Information System**]: Miscellaneous

Keywords

Recommendation; User Modeling; Hybrid Methods

1. INTRODUCTION

One common thread in recommender systems research is to combine recommendation methods as a strategy to improve performance [7]. Existing recommendation methods have strengths and weaknesses, and researchers proposed a large number of combination strategies. In most cases, Collaborative Filtering (CF) methods, which correlate user ratings with items, are combined to Content-based (CB) methods, which correlate user ratings with item attributes [1, 8]. While CF methods assume that users with similar consumption history would share common interests, CB methods conjecture that each user exhibits a systematic preference correlated with some item attributes [18]. Such preferences may stem from usual content-related data, such as genres or movie actors, but they may also come from metadata or consumption attributes that are not usually handled as "content", such as popularity, recency, or similarity to other consumed items.

This work evaluates whether those non-content preference attributes are properly captured by CF methods, and, wherever they are not exploited, how they may be incorporated into CB models to produce better hybrid recommenders. First, aiming to evaluate to what extent a non-content preference attribute is captured by CF models, we define the *preference mismatching* metric that quantifies how the recommendations provided by a given CF match the user previous consumptions w.r.t. the values along such attribute. Further, we present a characterization methodology to evaluate the preference mismatching metric for any CF method in real domains. Finally, we implement a method to build CB models by using non-content attribute information and combine these new models with existing CF methods to produce better recommendations.

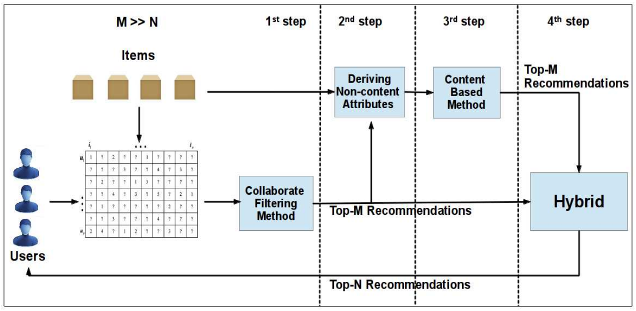

Figure 1: Our hybrid method

Figure 1 depicts our four-step hybrid method for the Top-N recommendation task. In the first step, we execute a given CF method, providing an ordered recommendation list L of size $M \gg N$. We then derive, in the second step, item attributes based on non-content attributes such as popularity, recency, and similarity. These derived attributes conform a vector space V where each item from L is represented. In the third step, we build a CB model for each user u considering attribute space V and determine a new score for each item from L, based on the preference mismatching metric. Finally, in the fourth step, we combine the CF score with the CB score, re-ranking the recommendation list L, and issue the N best ranked items as a recommendation to user u. It is worth emphasizing that our method differs from previous efforts by considering attribute data different from content ones, specifically data related to aggregate consumption (popularity), metadata (recency), and data based on individual consumption (similarity) as attributes that are mod-

[*]This work was partially supported by CNPq, CAPES, Fapemig, and INWEB.

RecSys'13, October 12–16, 2013, Hong Kong, China.
Copyright 2013 ACM 978-1-4503-2409-0/13/10 ...$15.00.
http://dx.doi.org/10.1145/2507157.2507179.

eled and exploited from a CB perspective and integrated into a hybrid recommender.

We employ a somewhat unusual approach to compute our CB scores, designed to reflect the fact that CF recommendations may or may not adequately represent user's preferences in the dimensions we model. Instead of directly computing a CB score, and then the hybrid recommender mix the CF and CB scores, we compute a CB delta score, a CB score that already has built into it a reflection of the degree to which the CF recommendation reflects the modeled user preference. Hence, if a highly-ranked item already reflects the user's preferences in non-content attributes, the CB delta will leave this item where it is in the recommendation list. But if an item is over or under recommended relative to the preference dimensions, the CB recommendation may move it down or up as is appropriate.

We evaluate our method in real domains as follows. We first derive three non-content preference attributes that span various domains: popularity, similarity and recency. Popularity refers to the percentage of distinct users who have consumed each item. Similarity is a measurement of how similar the items consumed by the users are, measured using the pairwise cosine distance of the item consumption vectors. Finally, recency refers to how long an item has been consumed in a domain. Then, we instantiate the hybrid method, modeling the non-content preference of users through a multivariate Gaussian and linear function that combines CF and CB scores. We evaluate it experimentally using five distinct real data collections derived from *Netflix*, *LastFm*, *MovieLens* and *Echo Nest* systems and six well known representative CF techniques implemented and distributed by the MyMediaLite project[9]. Besides demonstrating the existence of significant preference mismatching in all experimental configurations (data collections and techniques), the results also evince that major gains, regarding precision, novelty and diversity simultaneously, are achieved by exploiting non-content preference attributes. We observe improvements that range from 15% to 100% on all datasets. Further, we found that our hybrid method is able to improve the recommendations for more than 40% of the users who have room for improvement in a Top-500 CF recommendation.

In summary, we highlight as a main contribution of this work the modeling and the use of information that, so far, has not been completely exploited by RSs. Our results indicate a new and relevant research direction in recommender systems, through which significant enhancements may be achieved. Finally, to the best of our knowledge, our work is the first effort that effectively exploits non-content preference attributes in RSs.

2. RELATED WORK

The accurate modeling of the user's preferences is a core task for recommender systems (RSs) [18, 16]. Despite the number of efforts on this direction, it still represents a complex task due to distinct challenges [1, 11]. First, individual user's preferences do not follow strict and easily predictable patterns. Second, different users exhibit distinct preferences. Finally, we could mention the lack of enough information about user consumption in several domains [10]. Aiming to overcome individual limitations of RSs for addressing these challenges, an increasing number of researches combine existing methods in order to provide more robust and effective models, defining the so-called hybrid methods.

In most of the cases, the Collaborative Filtering (CF) methods are combined with Content-based (CB) ones [1]. CF methods model dependencies between user ratings and items by using consumption information, assuming that users with similar consumption history share common interests for sake of recommendation [8]. Many of these methods are inspired by machine learning algorithms, such as nearest neighbors [19], neural network classifiers [5], induction rule learning [3], Bayesian networks [6] and latent factor models [12], among others. On the other hand, CB methods model dependencies between user ratings and item attributes, assuming that each user exhibits a systematic preference related to one or more content attributes from the item, such as price, author or genre of books, for instance [18]. As these methods often come from information retrieval and filtering, many of them are based on traditional approaches of this area [14]. By combining CF with CB, hybrid methods are able to attenuate the weaknesses of both while exploiting simultaneously their strengths for recommendation. In this sense, for instance, [13] presents a hybrid approach that incorporates multidimensional clustering into collaborative filtering recommendation models. Also, in [17] it was employed an evolutionary search for hybrid models following the Strength Pareto approach, which identifies hybrid models that are in the Pareto frontier. Burke [7] presents a broad survey on hybrid methods.

Despite all advances in RSs, mainly on hybrid methods, we believe that there are improvement opportunities to be exploited, since even state-of-the-art methods are not able to provide good recommendations in different real scenarios [10]. This work aims to exploit one of these opportunities. We assume that systematic preferences of users may also be correlated to non-content attributes that refer to metadata or consumption. As such correlation is not modeled explicitly by CFs, we believe that they are under captured in practice. Thus, we model explicitly these correlations through CB models that are combined with traditional CFs, defining a new hybrid method that exploits complementary information available in the consumption data. Indeed, results related to our method demonstrate the usefulness of this information for enhancing recommendation. Further, we did not identify in the literature efforts on explicitly modeling and exploiting non-content attributes as we do.

3. NON-CONTENT PREFERENCE MISMATCHING

3.1 Non-Content Preference Attributes

A formal description of the Preference Mismatching requires, first, to define non-content preference attributes. A non-content preference attribute is derived from previous consumption data and quantifies a criterion by which an item might have been chosen previously. For instance, how long an item has been consumed in a domain or the item's popularity are non-content preference attributes available in almost all domains. Specifically, such attributes fulfill three requirements:

1. The consumption data are enough for allowing the computation of the attribute values;
2. There is a function that maps the represented criterion to a numeric spectrum of values, so that it is possible to assign a value to each item in the population;
3. The per user consumption must be related to a short range of values along the spectrum. In practice, we focus on attributes that may be exploited for sake of prediction. For instance, the last moment in which an item was consumed in a domain would not be relevant whether the user consumption disregards this information.

3.2 Preference Mismatching

Aiming to measure how well a specific non-content preference attribute is captured by RSs, we describe each item by D non-content attributes and define the statistical measure *Expected Value*

$(E(D_i))$ for any subset of items along each attribute D_i. In practice, such expected value can be approximated by the mean value observed in training data. Thus, we mathematically define user's preference in recommendation domains through Definition 1.

DEFINITION 1. *The* ***user's preference*** *is a D-dimensional vector that quantifies, for each attribute D_i, the relative difference between the expected value $E(D_i|S)$, given the subset S of items consumed by a specific user, and the expected value $E(D_i|I)$ given the entire item set I.*

More formally, let C_{u_k} be the set of items consumed by user u_k, during a given period of time. Also, let $E(D_i|C_{u_k})$ be the expected value of the set C_{u_k} for each $D_i{\in}D$. Thus, the preference $Preference[u_k, D_i]$ is given by Equation 1.

$$Preference[u_k, D_i]{=}\begin{cases} \frac{E(D_i|C_{u_k})-E(D_i)}{E(D_i)}, \text{ if } E(D_i){\neq}0 \\ \lim\limits_{E(D_i)\to 0} \frac{E(D_i|C_{u_k})}{E(D_i)}, \text{ otherwise} \end{cases} \quad (1)$$

We assume that consumption is not random and users present a systematic preference about items from a specific range of values for distinct non-content attributes. Therefore, explicitly considering such non-content preference allows us to refine the identification of user interests. For instance, the information that a specific user's preference is towards old and unpopular items allows selecting a subset of items that better suit his/her interests.

Analogously, we define the per user recommendation's non-content description as the relative difference between the expected value $E(D_i|R)$ observed in the subset R of items recommended to a specific user and the expected value $E(D_i|I)$ of the entire item set I. Let R_{a_n,u_k} be the set of items recommended by a method a_n to a specific user u_k. Further, let $E(D_i|R_{a_n,u_k})$ be the expected value of R_{a_n,u_k} along each $D_i{\in}D$. Thus, the description $Desc[a_n(u_k), D_i]$ of R_{a_n,u_k} is given by Equation 2.

$$Desc[a_n(u_k), D_i]{=}\begin{cases} \frac{E(D_i|R_{a_n,u_k})-E(D_i)}{E(D_i)}, \text{ if } E(D_i){\neq}0 \\ \lim\limits_{E(D_i)\to 0} \frac{E(D_i|R_{a_n,u_k})}{E(D_i)}, \text{ otherwise} \end{cases} \quad (2)$$

Such as expected for users, we assume that recommenders prioritize a specific range of values for each non-content attribute, since they are based on inductive premises that make some assertions about items or users. Thus, from this perspective, a relevant question concerns the match between user non-content preference and recommendation non-content attributes. Aiming to evaluate such match, we define a metric named **Preference Mismatching** that measures the difference between our mathematical definition of user's preference and the recommendation's non-content description. We assume that this difference can be calculated using different approaches and it is significant when its absolute value is higher than a minimum positive value ϵ that determines whether a recommendation description is similar to a user preference in a given domain. Therefore, a relevant hypothesis to be assessed is whether Preference Mismatching is usually significant in real domains.

4. ASSESSING PREFERENCE MISMATCHING EXISTENCE

Aiming to verify the existence of non-content preference mismatching in real domains, we present a characterization methodology that answers some crucial questions:

1. Is the user consumption associated with a short range of values for each attribute D_i?
2. What is the user's preference w.r.t. each attribute D_i?
3. Does the user consumption present high variability w.r.t. his/her individual preference for each attribute D_i?
4. What is the recommendation's non-content description w.r.t. each attribute D_i?
5. What is the preference mismatching w.r.t. each attribute D_i?

We conduct analyses related to each of these questions on a set of transactions T, which comprises the transactional history of users in a domain. Further, we divide T into two disjoint sets, a training set T_a and a test set T_e, such that $T{=}T_a{\cup}T_e$. In all steps, only the test set T_e is used for calculating the expected values and all other measures related to each non-content attribute. In turn, the training set T_a is used for assigning the values, for each attribute D_i, to each test item. For instance, the popularity inherent to each test item is defined as its prior popularity on T_a.

We evaluate the correlation between the user consumption and a given non-content attribute D_i through the *Normalized Standard Deviation* (NSD) defined for each user u_k, as described in Equation 3, where $\sigma(D_i|C_{u_k})$ denotes the standard deviation of values of D_i in the set of consumed items C_{u_k}. This metric assumes that a non-content attribute provides predictive relationship w.r.t user consumptions whether the dispersion of values observed for the consumed items is significantly smaller than the dispersion observed in the whole spectrum of values. Thus, the smaller the NSD, the more the user consumption is correlated to a subset of values.

$$NSD(u_k){=}\frac{\sigma(D_i|C_{u_k})}{|max(D_i){-}min(D_i)|} \quad (3)$$

Considering the assessments of non-content preference in real domains, we measure, for each user u_k, his/her $Preference[u_k, D_i]$ for each attribute D_i, as described by Equation 1. Besides preference, the consumption variability exhibited by each user is also relevant. Predicting consumption of attributes that present small variability w.r.t. its values tends to be easier, since the consumption becomes similar to the user preference. Conversely, when this variability is more pronounced, information about expected values becomes less useful and predicting future consumption becomes more challenging. We measure the contribution of each attribute D_i to the user consumption variability through the *Relative Standard Deviation* (RSD) defined for each user u_k, as presented in Equation 4. The higher the RSD, the higher the variability that D_i brings to u_k consumption.

$$RSD(u_k){=}\frac{\sigma(D_i|C_{u_k})}{E(D_i|C_{u_k})} \quad (4)$$

Similarly to the measurements of user preferences, the per user recommendation's non-content description is measured through the $Desc[a_n(u_k), D_i]$, described by Equation 2. Finally, we evaluate the Preference Mismatching for each user u_k as the difference between $Desc[a_n(u_k), D_i]$ and $Preference[u_k, D_i]$. Thus, we can verify whether CF models fail to incorporate accurate information about non-content preferences. The relevance of such analysis is that whenever the per user recommendation's non-content description differs significantly from the user consumption, the user interests are not satisfied, affecting the quality of the recommendations.

5. EXPLOITING PREFERENCE MISMATCHING ON RECOMMENDERS

In this section, we present our hybrid recommendation method for the Top-N recommendation task that combines rating information, assigned by traditional CF models to items, with the score defined by CB models that represent how well an item matches the user non-content preference. In this sense, we assume that the smaller the preference mismatching value, the better an item matches the user non-content preference. Therefore, explicitly approximating the recommendation non-content description to the user preference should mean a significant recommendation improvement. From this perspective, an item should be recommended to a specific user whenever, besides exhibiting a high rating, it has a high probability of matching this user non-content preference for a set of selected attributes.

Our method consists of four main steps. First, we execute a given CF method, such as *Matrix Factorization*, in order to obtain an initial list of M items deemed as relevant by the CF model, such that $M \gg N$. In the second step, we derive non-content attribute values and define a vector attribute space composed of these derived attributes. Then, we represent each item present in the CF recommendation list within this space by computing the item value along each attribute. In this paper we derive three attributes: **popularity**, **similarity** and **recency**. We selected these attributes based on some economic and social theories currently employed in RSs [2], which suggest that similarity, recency and popularity may be related to the user's taste. Further, previous studies have pointed out evidences of systematic trends along these attributes, reinforcing their relevance for this study. For instance, in [21], the authors argue that RSs are more apt to recommend popular items, while recommending unpopular ones remains a challenge.

We now define more formally each non-content attribute. Popularity refers to the receptivity of items in a domain, with respect to the desire of consumption. We measure its values as the percentage of distinct users who have consumed each item, regardless when, in a data sample. Similarity measures to what extent the items consumed by the user are similar to each other, using the pairwise cosine distance of the item consumption vectors. A user's preference for similarity is then computed as the average of the similarity scores of each consumed item with all other items consumed for that user. When assessing an item that is candidate for recommendation, its similarity score is the average of the similarity scores of that item with all already-consumed items for that user. Finally, recency refers to how long an item is available in a domain. We measure its values as the difference between a reference timestamp and the timestamp when the item was first consumed in a domain. By these definitions, it is straightforward that the selected non-content attributes fulfill the two first requirements discussed in Section 3.1. Since the third requirement is domain dependent, we evaluate it in the case study section.

In the third step, we define a CB model for preference on the vector space using a multivariate Gaussian due to its computational simplicity and the lack of evidences for adopting a more specific model. Thus, the non-content preference of each user u_k is a function $\mathcal{N}(\mu, \sigma^2)_{u_k}$ derived from the user non-content preference information. We define the average of the values, along each attribute, of all items already consumed by u_k as the mean μ, and the covariance matrix σ^2 is derived from u_k consumption history. Then, for each item t_j issued by the CF method in the first step, we define a new score that quantifies the preference mismatching between the item representation and the user preference model. In this case, such score is simply the probability defined by the function $\mathcal{N}(\mu, \sigma^2)_{u_k}$ at the point defined by the vector that represents t_j. The adoption of a probabilistic perspective for measuring the preference mismatching in this case stems from the need of models to take into account distinct attributes simultaneously and to capture both the user non-content preference and the variability around this preference. Differently from the characterization methodology, where we calculate the preference mismatching through an Euclidean perspective due to the goal of just measuring individually the mismatching along each attribute, we used a more robust perspective of analysis.

The last step combines the rating information provided by the CF method with this probabilistic score, generating a final score used for re-rank the recommendations. Among the possible combination strategies, we choose a simple linear combination between ratings and probabilities for defining the final score of each item t_j, such as presented in Equation 5.

$$Score(u_k, t_j) = \alpha \cdot \frac{R_{a_n,u_k}(t_j)}{\max R_{a_n,u_k}(*)} + (1-\alpha) \cdot \frac{G_{u_k}(t_j)}{\max G_{u_k}(*)} \quad (5)$$

where $R_{a_n,u_k}(t_j)$ denotes the rating assigned by a_n to t_j considering the target user u_k; G_{u_k} refers to the Gaussian distribution that models the user non-content preference and variabilities in the training set and α represents a weighting factor in the linear combination. In this work we just perform an exhaustive evaluation of several α values between 0 and 1, aiming to evaluate the relevance of the non-content preference information on this combination. Also, given the complexity of evaluating individual α values for each user, we adopted a single global α for all users, although it is expected that distinct users require different combination weights. Furthermore, we normalize each rating $R_{a_n,u_k}(t_j)$ and probability $G_{u_k}(t_j)$, since they vary on distinct scale of values.

Finally, we highlight that such hybrid method can be easily incorporated to the traditional recommendation process, regardless the domain or adopted CF. Whenever it is possible to identify any significant non-content attribute that users follow, we have an approach to incorporate it explicitly into the recommendations. Also, we can apply distinct strategies for building preference models, calculating preference mismatching and combining CF and CB scores.

6. CASE STUDIES

In this section, we assess the existence of the non-content preference mismatching and practical usefulness for recommendation in real domains. We start by presenting the data collections used in our experiments. Next, we briefly describe the evaluated CF methods. Then, we discuss the results related to our characterization methodology. Finally, we present the results related to the hybrid method that exploits the non-content preference.

6.1 Datasets

We perform the empirical evaluations considering five distinct real data collections. The first one is the well-known Netflix dataset from the movie domain. *Netflix* (http://www.netflix.com) is an online rental movie service that made available, for research purposes on recommendation, a database with information about its movies and users. As the second collection, we employ a sample from the *Last.fm* system (http://www.last.fm/), which is a UK-based Internet radio and music community website. This sample was collected through an API provided by *Last.fm* (http://www.last.fm/api). Our third and fourth datasets (ML-1M and ML-10M) comprise rating data samples from MovieLens (http://movielens.umn.edu), gathered and made available for research purposes by GroupLens Research. Finally, as the fifth collection, we choose a random sample from the Million Song Dataset (http://labrosa.ee.columbia.edu/millionsong/tasteprofile) [4], made available recently for research purposes on recommendation. Table 1 summarizes the main features of each evaluated dataset. All selected non-content attributes were evaluated in these five datasets, except the recency attribute that could not be evaluated in the Million dataset, since it does not provide temporal information about user actions. Also, we should mention that, for the calculation of each non-content attribute, rating based datasets were transformed into consumption data simply by considering all ratings as consumptions, disregarding the rating.

Table 1: Dataset information.

	Netflix	LastFm	ML-1M	ML-10M	Million
# Users	480,189	35,000	6,000	72,000	200,000
# Items	17,770	4 million	4,000	10,000	348,360
# Actions	100 million	85 million	1 million	10 million	19 million
# Time	310 weeks	281 weeks	149 weeks	671 weeks	-
Type	rating	play count	rating	rating	play count
Domain	movies	songs	movies	movies	songs

6.2 Evaluated CF Methods

Our analyses also take into account six representative CF techniques, both memory-based and model-based, for the Top-N recommending task. Specifically, the set A of evaluated methods comprises the algorithms *Matrix Factorization* (MF), *Latent Feature Log Linear Model* (LF), *Biased Matrix Factorization* (BMF), *SVDPlusPlus* (SVD) implemented and distributed by the MyMediaLite project [9]. For simplicity of analysis, we use the default parameters of each algorithm in the library on all evaluations. Furthermore, since the memory-based implementations of MyMediaLite are not able to handle the analyzed datasets, we implemented our versions of the traditional algorithms *UserKNN* and *ItemKNN* using the Cosine measure as similarity function, such as presented in [1]. Also, for both algorithms, we incorporate the sample bias regularization with the original parameters used in MyMediaLite and 80 as the maximum number of neighbors. Our experiments were performed in octa-core machines with 96 *GB* of RAM. However, we should mention that these machines were not able to run LF, SVD and ItemKNN methods on LastFm and LF method on our MillionSongs data sample, due the inability of these methods to scale to huge volumes of data.

6.3 Assessments of Preference Mismatching Existence

Starting our analyses by the measurements of consumption correlation, we plot a *Complementary Cumulative Distribution Function* (*CCDF*) of the NSD values found for each user. This distribution shows that, in general, users consume items belonging to a restricted range of values along each attribute. In all datasets, we observe that more than 80% of the users exhibit a normalized standard deviation smaller than 25%, 30% and 15% for popularity, recency and similarity, respectively. It means that the variability of consumption exhibited by each user usually relies on a range of values smaller than one quarter of the whole spectrum. Hence, despite not being sufficient for determining accurate preferences for each user, values in each of these non-content attributes may help to filter out irrelevant items.

By taking into account the user preference analysis, we plotted a *CCDF* of the $Preference[u_k, D_i]$ values found for each user in our data collections, such as presented by Figure 2. We observe that users exhibit distinct preference for each attribute. Further, the absence of gaps in these plots evinces that there is no predominant preference in the evaluated attributes. We observe low probabilities related even to zero, marked as dashed lines in the plots, although there is a concentration of preferences in a range near to zero (-0.5 and 0.5) in almost all cases. Basically, it means that user non-content preferences mostly deviate from a single and global expected value $E(H)$ half of the $E(H)$ value, for each attribute. Therefore, a single expected value is not enough for describing accurately all users. Finally, we point out that users exhibit a slightly higher interest towards more popular, similar and recent items than the expected in almost all datasets, since the probability of positive preference values along popularity and similarity is 60% and negative preference values along recency is 65%.

By plotting a *CCDF* of the RSD values found for each user, we evince large variabilities in all evaluated datasets, for the three selected attributes, as shown by Figure 3. Through these plots we observe that variability values higher than 50% are very likely in almost all datasets and attributes. That is, the length of the total variability observed for each user is larger than the expected value estimated from his/her consumption history. Thus, besides presenting distinct preference values, we observe that users also consume a range of items with different characteristics w.r.t. each attribute. Based on these results, we conclude that users are not strongly tied to their individual preference, presenting high variability of consumption in all evaluated collections.

In order to assess the recommendation non-content description values in the evaluated datasets, we plot a *CCDF* of $Desc[a_n(u_k), D_i]$ for each user u_k. Similarly to users, we observe that RSs provide distinct non-content descriptions, which also vary according to each dataset. Starting by recency, despite presenting distinct descriptions, we observe almost consensual behaviors among the six evaluated RSs. Most of them prioritize recent items, presenting over than 60% of probability for negative description values. For popularity and similarity, we observe a more diversified scenario. The same methods exhibit distinct behaviors on different datasets. For instance, whereas LF and SVD present positive popularity and similarity description values in ML-1M and ML-10, they exhibit negative ones in Netflix. Further, most of these non-content description values lie between -0.5 and 1.0 for both attributes, demonstrating a high diversity of descriptions. Also, the results show an unexpected behavior for UserKNN and ItemKNN w.r.t. popularity and similarity. Differently from previously stated [15], KNN-based methods prioritize items less popular and similar than those usually consumed by each user, exhibiting negative popularity and similarity description values. Such divergence stems from the fact that the consolidation of neighborhoods is heavily based on more popular and similar items, since similarity is usually defined over a consumption intersection between user transactions. However, the items actually recommended, which are outside of this intersection, tend to be less popular and similar.

Finally, analyses on the preference mismatching demonstrate that all evaluated methods provide recommendations that systematically deviate from the user preferences. Figure 4 shows a *CCDF* of the difference $Desc[a_n(u_k), D_i] - Preference[u_k, D_i]$ defined for each user u_k. We observe a high concentration of difference values near to zero within ranges of ± 1, ± 0.5 and ± 0.05 for popularity, similarity and recency, respectively. Thus, an ϵ value of 0.33, for instance, would be enough for defining a significant preference mismatching along similarity for more than half of users in almost all datasets, methods and attributes. Indeed, these are expressive differences when we take into account the expected values. For example, for a user who exhibits a similarity preference of 0.30, such difference means that RSs usually recommend items with non-content description from 0.20 to 0.40. Considering each attribute individually, through recency we observe that, besides presenting description values towards recent items (i.e., negative values), most RSs present recommendation descriptions stronger than the user preferences, recommending to them items more recent than they usually consume. For popularity and similarity, RSs present diversified behaviors. Whereas positive description values are observed in some datasets, negative ones, related to the same RSs, are observed in other datasets. We believe that these behaviors result from inherent characteristics of each dataset and a deeper analysis in this direction would be required.

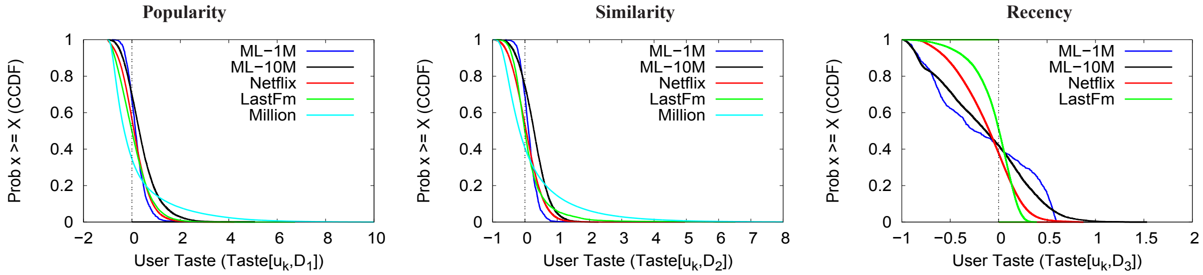

Figure 2: Analysis of user preferences.

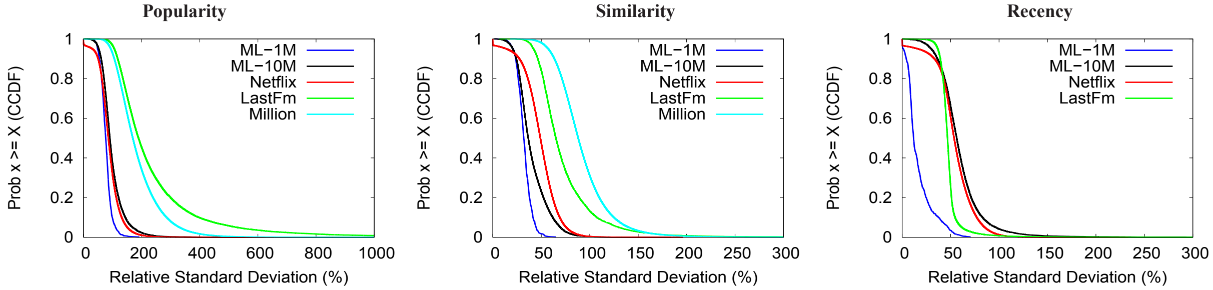

Figure 3: Distributions of relative standard deviations of user consumptions.

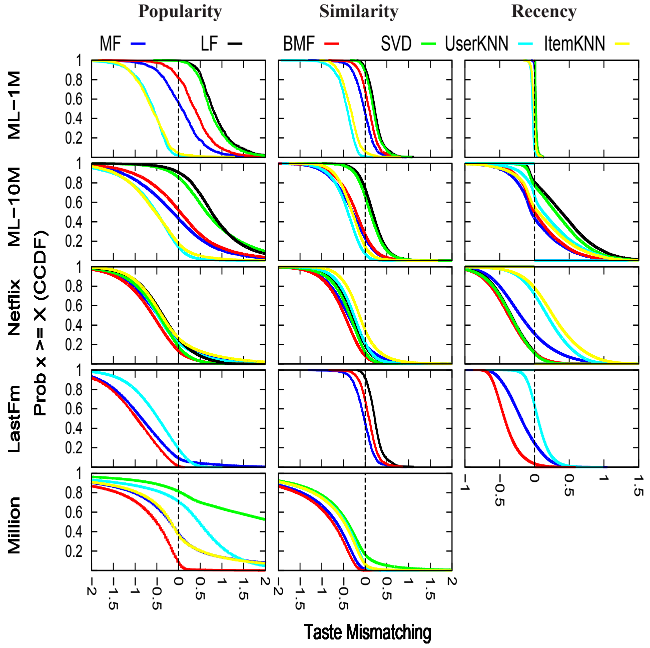

Figure 4: Analysis of preference mismatching.

6.4 Assessments of Preference Mismatching Exploitation

Besides demonstrating the existence of significant non-content preference mismatching in real domains, it is of paramount relevance verifying the utility of reducing this mismatching towards better recommendations. In this sense, we evaluate our hybrid method. Since a proper *n-fold cross validation* design would require a careful design in temporally ordered data and demand huge execution time for the evaluated datasets and algorithms, the following analyses employs a traditional training (70%) / test (30%) partition. Further, aiming at a broader notion of quality in recommendation, our analyses take into account three distinct quality dimensions for a Top-50 recommendation task: accuracy, novelty and diversity. Assessments on accuracy are based on the classical $Precision@50$ and precision is measured by counting the number of distinct items of the Top-50 recommendation that appears in the per user test set. We measure novelty and diversity through a formal framework of analysis presented in [20]. More specifically, we use the EPC_rank, and the EILD for measuring novelty and diversity, respectively, considering in both cases the discount function ($disc(K)$) equals to 0.85^{k-1}, Pearson correlation as similarity distance measure and relevance aware recommendations [20]. Also, we set the parameter M given as input for our hybrid method to 500, aiming to exploit significantly larger lists of items than the final recommendation list (ten times larger in this case) while keeping computationally feasible the experimentation. Finally, we point out that our strategy of analysis is based on contrasting the results of each original CF method a_n against the results of our hybrid model when performed with a_n. Our primary goal is to identify the relevant of non-content preference attributes for improving traditional CFs, rather than contrasting it against other hybrid methods.

We start our analyses by investigating the individual usefulness of each selected non-content attribute, for providing better recommendations. Figure 5 shows the gains and loses of $Precision@50$ when building a probabilistic model for each attribute individually and using distinct combination weights. We observe expressive gains when exploiting the popularity attribute in MF, BMF, UserKNN and ItemKNN in all datasets. However, the methods LF and SVD, which exhibited the highest popularity preference mismatching values in Figure 4, could not be improved through this attribute. As they exhibit such a strong deviation towards popular items, reducing preference mismatching among the 500 recommended items was not enough to improve the results. Taking into account similarity and recency, we could not improve the CF results in most of the cases. In summary, it shows that our hybrid method is not able to effectively exploit alone each of these attributes in order to improve recommendations.

An immediate question is what happens when we take into account the three selected attributes simultaneously. Figure 6 answers this question regarding accuracy, novelty and diversity. We observe that, besides even higher gains in terms of accuracy in almost all CF algorithms and datasets, we also achieve simultaneous gains w.r.t. novelty and diversity. In general, the most expressive gains

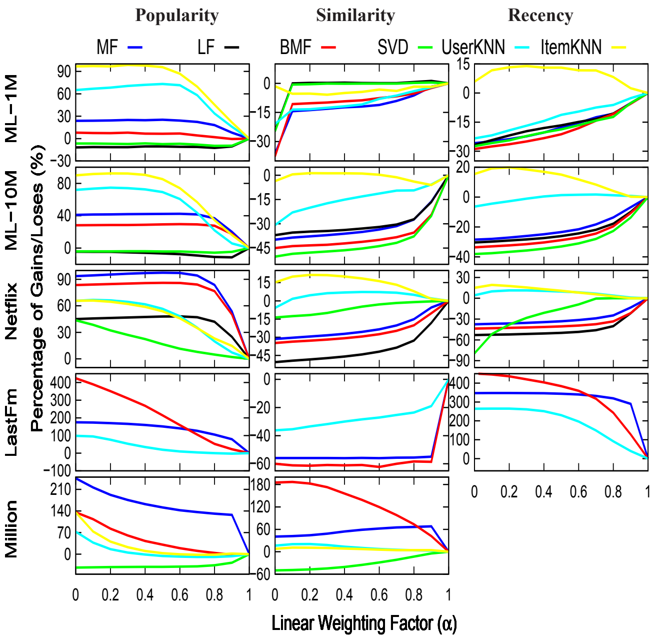

Figure 5: Analysis of $Precision@50$ gains|loses by exploiting, individually, popularity, similarity and recency.

were observed in CFs with the worse performance in each dataset. For instance, we observed gains over 200% on LastFm and Million datasets. In these cases, the original CF results were actually not significant. However, among the Top-500, the CFs were able to rescue several relevant items for each user and the non-content preference information was enough for identifying these items. On the other hand, gains around 10% were consistently related to CFs focused on suggesting items more popular, similar and older than the user expected interest (i.e., LF and SVD). Therefore, although they could achieve high accuracy rates, several recommended items not necessarily suit the user non-content preference. This fact explains why the gains in these cases are not as expressive as for the other methods. As the lists provided by such CF exhibit non-content description far from the non-content preference of each user, the suggested items do not necessarily present characteristics close to the user interests. Thus, items closer to the user preference in the Top-500 are still far from the actual user non-content preference.

Besides verifying the strength of the improvements provided by exploiting non-content preference in RSs, it is also important to investigate how often these gains happen. That is, we also evaluated the percentage of users in each database for whom our hybrid method was able to produce any improvement in a Top-N recommendation, in terms of accuracy. As our method processes a prior recommendation list of size $M \gg N$ provided by a CF method for each user, such percentage is limited by the percentage of these prior lists that have more relevant items than those present among its N first items. Figure 7 presents the percentage of possible improvements as the number of distinct users for whom our hybrid method enhances, divided by the number of users for whom the recommendation list provided by each CF could be improved. Our method, even adopting a global linear combination weight, is able to improve recommendation for more than 40% of these users in most of the cases. Further, the percentage of user for whom our method produced losses was at most 10% in all datasets.

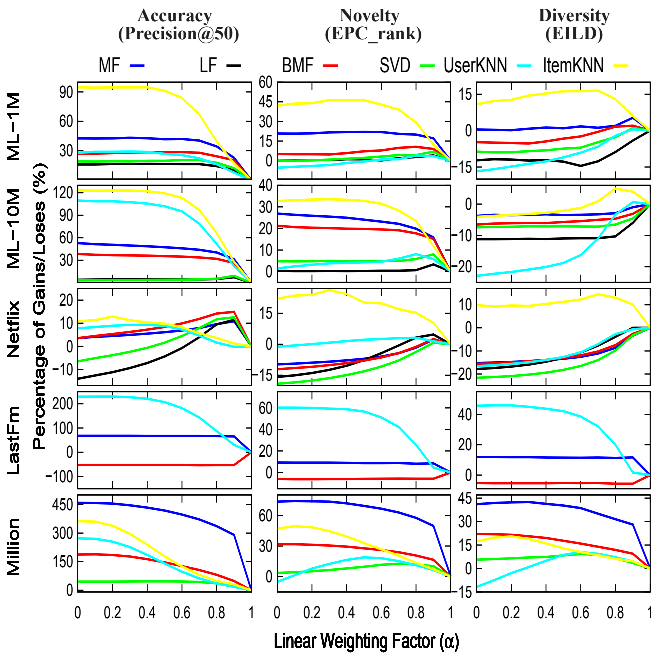

Figure 6: Analysis of quality gains|loses by exploiting, simultaneously, popularity, similarity and recency.

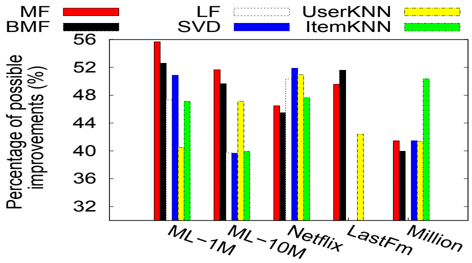

Figure 7: Percentage of users for whom the non-content preference information produced any improvement.

In summary, our hybrid method allowed us to verify the relevance of the user non-content preference in practical scenarios. Through this information, we are able to provide expressive gains in terms of accuracy, novelty and diversity in six major current CFs in the Top-50 recommendation task. We explain these gains by the fact that non-content preference information is able to filter out items that seem to suit the user preferences, but mismatch non-content characteristics from the items consumed by each user. Aiming to evince the existence of such mismatch items among the sorted Top-N list originally recommended by each CF, we calculate the non-content preference mismatching, such as performed in our methodology, but considering now each rank in this list. For sake of brevity, we show the mean of this deviation per rank among all users for one dataset (ML), such as presented in Figure 8, although the same behavior was observed in all other collections. We observe that the preference mismatching varies significantly among the ranks, not presenting any monotonic behavior. It reinforces that CF recommendations do not capture the systematic preference existing along each non-content attribute.

Additionally, we evaluate to what extent Gaussian functions are appropriate for modeling the user non-content preferences for the selected attributes. By plotting the probability distribution of the mean expected value of all users, along each attribute, and its standard deviation, we found that, indeed, the global behavior in our domains present a Gaussian shape. However, these distributions point out two main issues. First, users exhibit behavior significantly distinct from each other, demonstrated by high standard deviations in all points. Hence, distinct users would require, besides different parameters, different model functions. For instance, while a set of

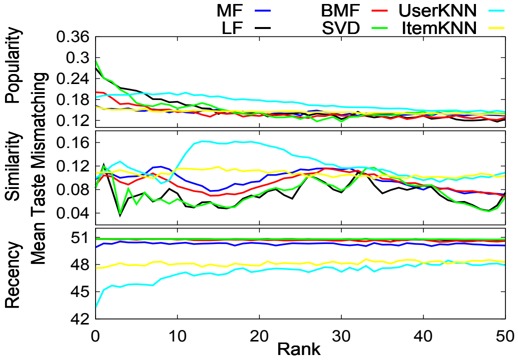

Figure 8: Analysis of mean preference mismatching per rank in CF recommendation lists.

users exhibit a Gaussian-like behavior w.r.t. the preference, others would present a power law like. In this case, non-parametric solutions should be applied. Second, we observe that the evaluated attributes do not vary in the same way, requiring distinct models. While recency seems much more an exponential function, popularity and similarity present a Gaussian shape. Therefore, besides all gains achieved by our simple method, these issues point out a room for even more improvements.

7. CONCLUSIONS AND FUTURE WORK

This work evaluates the use of attributes thought of as "non-content" towards better recommendations. We define as non-content attributes those related to metadata or consumption information. Specifically, we evaluate the attributes of popularity, similarity and recency derived for each item from consumption data. Besides demonstrating that recommendations issued by current CF methods do not capture properly the user preference along these attributes in real domains, through a characterization methodology, we implement a hybrid method that exploits such attributes in practice. Our method defines a CB model for user non-content preferences through a multivariate Gaussian function. We use this model for defining a score for each item, previously issued by a CF method, which represents how well this item matches the user non-content preference model. Finally, we combine the original CF score with the CB score, re-ranking the CF recommendation list.

Empirical evaluations conducted on five distinct real datasets and six representative CFs allowed us to verify the relevance of the user non-content preference in practical scenarios. Indeed, we achieved expressive gains regarding accuracy, novelty and diversity simultaneously in a Top-50 recommendation task, which range from 15% to 100%. Further, we found that this simple method was able to improve the recommendations for more than 40% of the users who have room for improvements in a Top-500 CF recommendation.

Besides significant enhancements over CFs, the exploitation of non-content attributes brings many questions not answered by this work yet. For instance, we are not able to make strong claims about the best approach for minimizing the preference mismatching or even the necessary conditions on which the implemented method provides gains. Further, we are aware of some weaknesses related to implementation and evaluation decisions. An immediate future step is conducting a comparative analysis of our method against other hybrid ones. Designing more robust experiments that balance time demanding and statistical robustness of analysis is also desirable. In addition, our gains are limited by the Top-500 recommendation lists provided by each CF. Whether the items recommended by a given CF do not cover those with characteristics similar to the user non-content interests, our method will not be able to bring improvements. Finally, simplistic implementation decisions, such as a single global linear combination weight α, should be refined to handle the actual conditions of recommendation domains.

8. REFERENCES

[1] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE TKDE*, 17(6):734–749, 2005.

[2] C. Anderson. *The long tail*. Gramedia Pustaka Utama, 2006.

[3] C. Basu, H. Hirsh, and W. Cohen. Recommendation as classification: Using social and content-based information in recommendation. In *Proc. of the NCAI*, 1998.

[4] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere. The million song dataset. In *Proc. of the 12th ISMIR*, 2011.

[5] D. Billsus and M. Pazzani. Learning collaborative information filters. In *Proc. of the 15th ICML*, volume 54, page 48, 1998.

[6] J. Breese, D. Heckerman, and C. Kadie. Empirical analysis of predictive algorithms for collaborative filtering. *Learning*, 9:309–347, 1992.

[7] R. Burke. Hybrid web recommender systems. *The adaptive web*, pages 377–408, 2007.

[8] M. Ekstrand, J. Riedl, and J. Konstan. *Collaborative filtering recommender systems*. Now Publishers Inc, 2011.

[9] Z. Gantner, S. Rendle, C. Freudenthaler, and L. Schmidt-Thieme. Mymedialite: a free recommender system library. In *Proc. of the 5th ACM RecSys*, 2011.

[10] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, 22(1):5–53, Jan. 2004.

[11] D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich. *Recommender systems: an introduction*. Cambridge University Press, 2010.

[12] Y. Koren. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proc. of the 14th SIGKDD*, pages 426–434, NY, USA, 2008. ACM.

[13] X. Li and T. Murata. Multidimensional clustering based collaborative filtering approach for diversified recommendation. In *Proc. of the 7th IEEE ICCSE*, 2012.

[14] P. Lops, M. Gemmis, and G. Semeraro. Content-based recommender systems: State of the art and trends. *Recommender Systems Handbook*, pages 73–105, 2011.

[15] R. Rafter, M. O'Mahony, N. Hurley, and B. Smyth. What have the neighbours ever done for us? a collaborative filtering perspective. *17th UMAP*, pages 355–360, 2009.

[16] P. Resnick and H. Varian. Recommender systems. *Communications of the ACM*, 40(3):56–58, 1997.

[17] M. Ribeiro, A. Lacerda, A. Veloso, and N. Ziviani. Pareto-efficient hybridization for multi-objective recommender systems. In *6th ACM RecSys*, 2012.

[18] F. Ricci, L. Rokach, and B. Shapira. Introduction to recommender systems handbook. *Recommender Systems Handbook*, pages 1–35, 2011.

[19] J. Schafer, J. Konstan, and J. Riedi. Recommender systems in e-commerce. In *1st ACM EC*, pages 158–166. ACM, 1999.

[20] S. Vargas and P. Castells. Rank and relevance in novelty and diversity metrics for recommender systems. In *Proc. of the 5th ACM RecSys*, pages 109–116. ACM, 2011.

[21] H. Yin, B. Cui, J. Li, J. Yao, and C. Chen. Challenging the long tail recommendation. *VLDB Endowment*, 2012.