A User-Centric Diversity by Design Recommender System for the Movie Application Domain

Michele Zanitti
CMI, Aalborg University Copenhagen
Denmark
mzanit15@student.aau.dk

Sokol Kosta
CMI, Aalborg University Copenhagen
Denmark
sok@cmi.aau.dk

Jannick Sørensen
CMI, Aalborg University Copenhagen
Denmark
js@cmi.aau.dk

ABSTRACT

Recommender systems (RS) have seen widespread adoption across the Internet. However, by emphasizing personalization through the optimization of accuracy-focused metrics, *over-personalization* may emerge, with negative effects on the user experience. A countermeasure to the problem is to diversify recommendations. In this paper, we present a solution that addresses the problem in the context of a movie application domain. The solution enhances diversity on four related dimensions, namely global coverage, local coverage, novelty, and redundancy. The proposed solution is designed to diversify users profiles, modeled on categorical preferences, within the same group in the recommendation filtering. We evaluate our approach on the Movielens dataset and show that our algorithm yields better results compared to random selection distant neighbors and performs comparably to one of the current state of the art solutions.

KEYWORDS

Recommender systems; diversity recommendation; user modelling; user clustering; movie recommendation; personalization

ACM Reference Format:
Michele Zanitti, Sokol Kosta, and Jannick Sørensen. 2018. A User-Centric Diversity by Design Recommender System for the Movie Application Domain. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3184558.3191580

1 INTRODUCTION

By exploiting user's preferences, recommender systems (RS) filter out irrelevant options and select only a personalized subset of items. Moreover, RS aim to promote the discovery of content to leverage the long-tail distributed consumption. There is an increasing concern regarding the issue of content over-personalization. If recommendations only mirror individual preferences, the resulting over-personalization could impact negatively the user satisfaction [1]. Whilst accuracy is important for user satisfaction, it is merely one ingredient. One approach is to diversify the recommendations for the users so that they do not meet their preferences completely [2]: diversified recommendations come at the expense of being inaccurate, but could contribute in a better item discoverability and a multi-faceted interpretation of user intentions [3].

WWW '18 Companion, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3191580

Our work provides the following contributions: *i)* a user-centric diversification framework which identifies the dimensions on which diversity can be modeled for RS. *ii)* We construct a novel category-based user model. *iii)* We propose a recommendation approach which implements diversity by design for the movie application domain, with the goal of finding dissimilar users and recommend items from their preferred categories.

The paper is organized as follows: Section 2 presents diversity as a theoretical concept, and popular metrics evaluating the quality of a diversification algorithm. Section 3 provides a comprehensive overview of current diversification solutions. In Section 4, we describe our contributions: the proposed diversification framework and the recommendation approach with embedded diversity. In section 5 we implement the proposed approach and evaluate the diversification levels reached by finding dissimilar users. For the implementation, we prepare the data by using the Movielens dataset, containing the user ratings, in conjunction with the IMDb metadata of the rated movies, relevant for the movie application domain. We then compare the proposed approach to a baseline method based on random diversification of users and to a state of the art method. Lastly, in Section 6 we discuss the results of the experimental evaluation and provide our plans for future work.

2 BACKGROUND

The concept of diversity is often understood as *heterogeneity*. Several techniques are available to quantify the diversity of a set of elements, depending on the interpretation of the concept itself [4–6]. Stirling [6] identifies three properties related to diversity and proposes a general framework to analyze it by considering the categories into which elements can be classified:
Variety: *"is the number of categories into which system elements are apportioned"*, i.e., the number of categories present in a set, independently from the elements within each, is a signal of diversity.
Balance: *"is a function of the pattern of apportionment of elements across categories"*. This assesses the extent to which categories are equally represented, through the relative distribution of elements.
Disparity: *"refers to the manner and degree in which the elements may be distinguished"*. This property assesses the specificity of each category (i.e., how can be easily distinguished), determining the dissimilarity as a signal of diversity.
Several works have proposed evaluation metrics for diversity in RS.

The *Binomial Diversity (BD)* [7], includes two parameters to evaluate the diversity in order to minimize the occurrences of similar items and to maximize the recommended items range with respect to the user preferences and to the item catalogue: coverage and redundancy. For a list L of recommendations, BD is measured as:

$BD_L = coverage_L \cdot nonRedundancy_L$, where $coverage_L$ can be considered either locally, in terms of user experience, or globally, as the ability of the recommender to consider multiple item categories [8].

Ziegler et al. define the *Intra List Similarity (ILS)* as the aggregate similarity between pairs of items in a set so that the lower it is, the higher the diversity and vice versa [9]. For all items in a list L and a pair of items i and j, ILS_L is:

$$ILS_L = \frac{1}{2} \sum_{i \in L} \sum_{j \in L, j \neq i} sim(i, j) \tag{1}$$

Following Stirling's definition [6], we propose an adaptation of Ziegler's *ILS* that considers the similarity of a pair of categories in which the recommended items are apportioned, instead of the single items. In fact, this metric can consider the disparity contained in the recommendation list, measured by the similarity by the items' categories.

Murakami et al. [10] define the *unexpectedness level* as the difference of the recommended items to the expected (as obvious) set of recommendations. However, since this definition works on single items, we adapt it to consider the user profile in terms of item categories. In the rest of the paper, we define the level of unexpectedness for a recommendation set L, considering the expected recommendation categories C_u for user u, as follows:

$$unexpectedness_L = \frac{|C_u \setminus C_L|}{|C_u|} \tag{2}$$

where, the item categories of the recommended list are C_L, in which each item is apportioned, and C_u is similarly defined but for the items the user has already experienced.

Lastly, we formulate a metric for estimating the redundancy obtained with a list of recommendations for a given user as the amount of items falling in the same category. We claim that the category-based *ILS* can be treated as a redundancy metric if the similarity between categories is as follows: $sim(C_i, c) = \begin{cases} 1 & C_i = c \\ 0 & otherwise \end{cases}$. In this case, the proposed metric considers only the categories C_u covered by u, which are the expected ones, and for a recommendation list, the "good" redundancy can be estimated as follows:

$$redundancy_{L,u} = \frac{1}{2} \sum_{i \in L, C_i \in C_u \cap C_L} \sum_{c \in C_u} sim(C_i, c) \tag{3}$$

where the redundancy is the $categoryILS_L$ of each covered category with respect to the recommendations list.

3 RELATED WORK

RS literature distinguishes two paradigms of diversification methods depending on the level at which is achieved, namely diversity modeling and post-filtering approaches [11].

The former solutions aim to enhance the filtering step by combining diversification criteria prior to the extraction of a set of recommendations. Instead, post-filtering methods process the set of candidate items after the filtering step through re-ranking strategies to extract the subset that satisfy the specified diversification and quality criteria [12]. Castells et al. [8] provide a unified understanding of diversity with an extensive survey on evaluation metrics, but a limited overview of diversification methods. Hence, we refer to [13], which provides an updated survey of both methods and evaluation metrics.

Ziegler et al. *topic diversification* [9]: this technique is devised to *"balance and diversify personalized recommendations lists in order to reflect the user's complete spectrum of interests"* and falls into the post-filtering solutions as it takes recommended items in input to re-rank and extract a diversified subset of items. Interesting, but perhaps unsurprising, is the application of item similarity, a content-based metric, to diversify the item set.

Vargas et al. *binomial diversity* [7]: instead of using item similarity, Vargas et al. proposed a definition of diversity encompassing the genre coverage, genre redundancy and the recommendation list size awareness. Still, since it works by re-ranking an initial recommendation, it falls into post-filtering approaches. Here, it is interesting how the technique approaches diversity taking not only the similarity, but also coverage and redundancy into account: coverage is achieved through finding the genres present in a recommendation set, compared to all genres, considering also that users themselves have preferences over certain genres and thus, some are more relevant than others. Redundancy in turn is defined as the frequency of each genre in the item set. The resulting method thus aims to maximize coverage of genres considering the user preferences while decreasing redundant genres [7].

ClusDiv [14] uses clustering to group items in the catalogue from the explicit ratings rather than on item descriptions (although this is not a necessary prerequisite), and recommends items from different clusters. Compared to re-ranking methods such as [9], using item clusters resulted quicker and achieved similar diversification results. The authors employed k-Means as the clustering method to generate clusters which were subsequently used to create a users–to-clusters weights matrix. However, as it takes a precomputed list of recommendations, it is a post-filtering approach.

Neighbor Diversification [15] proposes to retrieve a set of diverse users, by using explicit ratings, to an active user; recommendations are then extracted from these distant neighbors. Diversity is evaluated also considering the catalogue coverage, the novelty and the accuracy. An unexpected (and perhaps serendipitous) finding is that the accuracy levels, in terms of precision and recall, as the user diversity threshold increases, do not drop and in some cases can also increase, thus suggesting that the trade-off between accuracy and diversity may hold when considering items, but other factors come into play when users are considered.

XploDiv [16] employs Stirling's definition to suggest that the balance affects the trade-off between relevance and diversity and a novel diversification method has been devised to deal with the trade-off and with the user's openness tendency (to explore novel items or to exploit her preferences). The parameters to control the two trade-offs are tunable and dynamically learned, to allow a fine grained control of exploitative or exploratory diversity. Unlike the previous method, XploDiv is devised as a post-filtering approach and requires a set of recommended items.

4 A NOVEL DIVERSIFICATION FRAMEWORK

Given that the metrics presented focus on specific aspects to evaluate the diversity of a recommendation list, we unify these aspects and propose a user-centric conceptual diversification framework

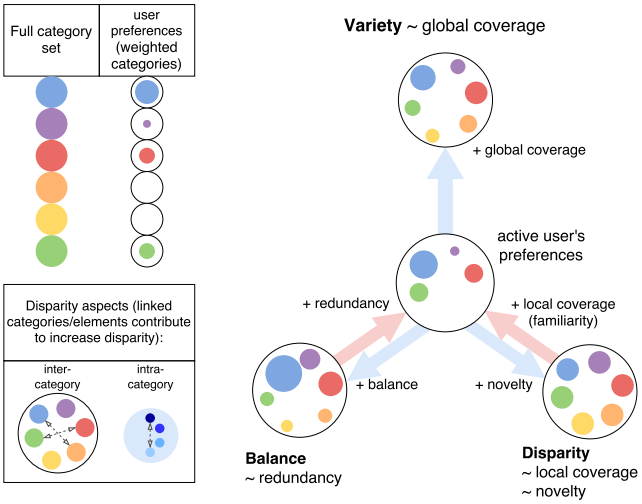

Figure 1: Proposed user-centric diversification framework: each small circle refers to the size of category-based preferences (small: low preference, see top-left corner); the three big circles stemming from the centre refer to the addressed diversity properties (global coverage, redundancy, local coverage and novelty).

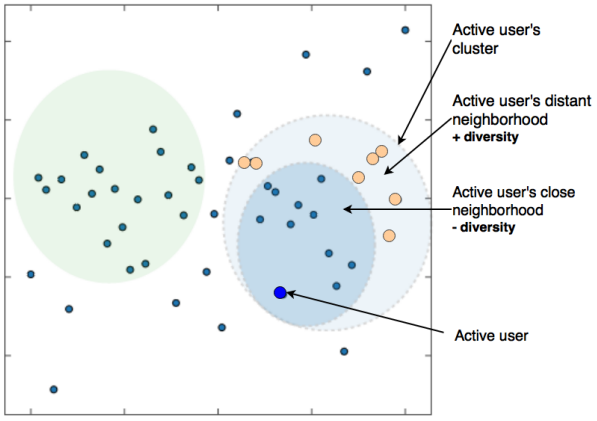

Figure 2: Proposed diversification approach: filtering distant neighbors (orange points) in the same cluster of users.

(Figure 1) built on top of four general dimensions where diversity can be controlled and evaluated for the individual user, namely: local coverage, global coverage, redundancy (Equation 3), and novelty (Equation 2). We adapt Stirling's definition [6] to accommodate the concept of users as mixtures of categories and that categories of items are weighted differently, since the concepts of personalization and user preference (either explicit or implicit) in RS conflict with how equally important the original definition treats the diversity of categories. In particular, user preferences would undoubtedly conflict with the property of balance, which instead assumes that "the more even the balance, the greater the diversity" [6]. So, a relaxed version of the diversity definition would allow to adjust the level of personalization for a particular user (in terms of balance), in relation to her preferences, while maintaining an acceptable level of heterogeneity thanks to the other properties (variety and disparity).

Since the majority of methods presented in Section 3 falls into the post-filtering category, we propose a diversification technique that is tightly coupled to the recommendation filtering, therefore can be formulated as a diversity modeling, in contrast to a post-filtering approach. In fact, post-filtering methods require an initial set of candidate recommendations as input to be diversified. This class of methods being dependent on the set, has the downside of requiring that the set is already diversified. This step is bypassed by the other class of approaches, which arguably allows a greater control of the diversification output. A high-level depict of the rationale behind the our diversity modeling approach is visualized in Figure 2. We suggest to utilize the categories of items as a way for modeling the user preferences and then, compute a list of recommendations by selecting distant neighbors for the active user. Moreover, in order to maintain a baseline level of accuracy in the neighbor filtering, we propose to group users together: from the users within the same cluster, only the distant ones are retrieved, while the nearest neighbors are filtered out. While retrieving the most similar users may be beneficial for the accuracy of the system, we argue that providing recommendations from distant users may increase the likelihood that the items are diverse and therefore, the chances for serendipitous encounters with novel items. Diversifying users would help to decrease the preference polarization by allowing more opportunities from the experience of different types of items. As a first idea, the approach would retrieve the distant neighbors from the same cluster, since we expect that the baseline within-group similarity can be a threshold for considering common preferences and therefore, the accuracy VS diversity trade-off, so that the users may not be completely dissimilar. It is important to recall that diversity is applied to the individual active user profile; hence, we consider five dimensions on which trade-offs between the user preferences and the other users' can appear:

Neighbor distance: The accuracy VS diversity trade-off is controlled by the pairwise diversity between the active user and the other users in the same group. As such, the diversity of the retrieved users should follow a parameter which can be tuned externally.

Variety of categories (global coverage): The number of categories that should appear in the recommendation list (without considering their disparity or the novelty to the user profile) depends on the length of the desired recommendation list and should be controlled to ensure that not too many or too few categories will appear. By controlling the variety of categories, the user will have the opportunity to navigate a list of recommendations with items belonging to many or few categories, hence, it can also be understood as the property of global coverage. Also, the variety implicitly controls the category disparity, as it seems reasonable that with low variety of categories comes a low disparity and vice-versa.

Disparity of categories (novelty and local coverage): The heterogeneity of the categories is the result of the novelty VS local coverage trade-off between the favorite categories for the active user and the distant neighbors'; therefore, it considers how many of the novels and how many of the covered categories should be used to select the items.

Variety of items: Similarly to the selection of the number of categories for the recommendation list, the variety of items considers how many items to show in the list. The variety of items is proportional to the variety of categories, as the number of categories increases, also the number of items should increase. The variety of items, in terms of list size, has been acknowledged to influence the coverage and redundancy of the recommended set [7]. In fact, it seems reasonable to generate recommendations from a number of categories proportional to the desired list length (i.e., for short lists, few categories and vice-versa); as such it is proposed to control the variety of categories accordingly.
Item recommendation list balance (redundancy): The last trade-off is controlled by the number of items belonging to the same category that should appear in the final list of recommendations. This trade-off is between balance and redundancy to control the visibility of each category; the bigger the redundancy, the more items belonging to the same category and vice-versa.

The designed technique takes inspiration from Yang et al. [15]. However, the aspects on which the proposed approach differs from [15] are as follows: i) In [15], the neighborhood is produced by maximizing the significance between the active user and the users in the neighborhood and subsequently selected. The proposed approach instead includes an additional step, the formation of groups of users, which considers a baseline similarity among the users within the same group. ii) In [15], the significance between two users is calculated to take into consideration the accuracy-diversity trade-off, controlled by a tunable parameter. The proposed approach extracts the users within the same group and, using a similar method, it controls the similarity VS diversity trade-off. iii) The technique illustrated in [15] utilizes solely the rating dataset, whereas we exploit both ratings and metadata. iv) In [15], the objective is to predict the ratings for unseen items for the active user, while here the ratings are the starting point to categorize user preferences. Nevertheless, there can be an additional step in the proposed approach to predict the relevance of a certain category for the active user, given the similarity between the user and the distant users and the preferences for their categories. v) However, the major difference lies in the categorization of user preferences (apart from the creation of groups of users), not included in [15].

4.1 Preprocessing Modules

We design the feature extraction modules, delegated to the transformation of user and item profiles into structures suitable for the recommendation procedure: item categorization, user category-based profile modeling and user group formation.

4.1.1 Item categories construction. The first step in the recommendation preprocessing methodology applies the vision of the diversity definition to compose a taxonomy of item categories (Figure 3). We propose to use the LSA methodology [17] as a preliminary step to the clustering analysis of the items to find the latent similarities based on the original extracted features: the feature occurrence matrix is computed, then TF-IDF (a weighting scheme operating on terms t contained in a document d, belonging to a collection of documents D) transformed to find a potentially unbiased subset of features and calculate their relevance scores for each item. Secondly, the item profiles dimensionality is inspected and SVD is applied to reduce the dataset dimensionality in order to extract latent features upon which items are compared. Clustering analysis is performed on the factorized item profiles, according to the criteria of similarity/distance metrics, clustering algorithm and number of desired clusters; specifically, the number of clusters is subordinated to a qualitative and quantitative analysis, in order to limit the creation of redundant clusters (i.e., having similar items in separate clusters), which can also be highly specialized and a potential cause of overfitting. As a result, the item categories are created and each item is labeled accordingly.

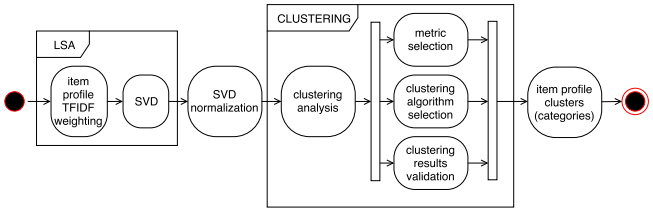

Figure 3: Item categories construction.

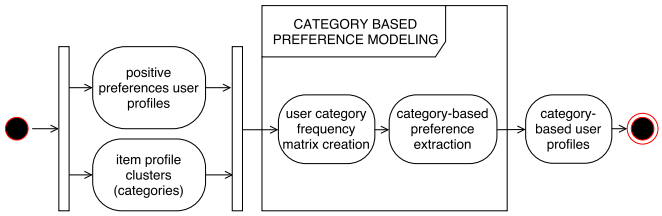

Figure 4: Modeling the category-based user profiles.

4.1.2 Category-based user preference modeling. models the user profiles according to the item categories. We extract the positive ratings for individual users according to their average ratings and individual thresholds τ, to remove the bias of different rating scales. This model assumes that the ratings are explicit, nevertheless, we argue that the same argument can also apply to implicit ratings. Figure 4 illustrates the process, which uses the output of the previous preprocessing modules as input to model user preferences and extracts the category-based user profiles. Two matrices are used for this purpose: R, containing the positive ratings of users u on items i, and the cluster matrix C, a boolean item i to cluster c association matrix. The profile of user u is constructed with the matrix $P_{u,c}$, where c is a cluster (or category) of items, and each element of the preference matrix is defined as follows: from the preliminary user profiles, the categories to which positively rated items are apportioned are extracted and the profiles are initially encoded with each category raw frequency (numerator of Eq. 4). Then, each user profile is divided by the number of categories experienced. By transforming the raw category frequencies into their proportion to the user profile, the differences towards users having experienced more categories and users having more focused interests may be more comparable using profile proportions, rather

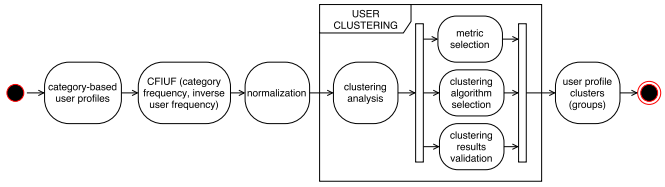

Figure 5: User clustering process.

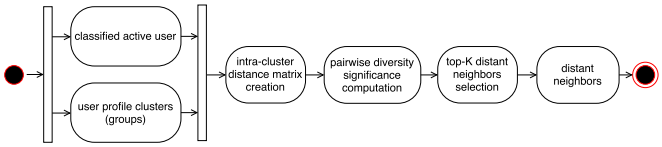

Figure 6: Diversification procedure in detail.

than raw frequencies and average ratings.

$$P_{u,c} = \frac{\sum_{i \in I_{u,c}} C_{i,c}}{\sum_{c \forall c \in C_u} \sum_{i \in I_{u,c}} C_{i,c}}, \text{where:} \quad (4)$$

- $I_{u\tau} = \{i | r_{u,i} \geqslant \tau, \forall r_{u,i} \in R_u\}$ is the set of items rated positively by u;
- R_u is the set of ratings for user u;
- $I_{u,c} = \{i | C_{i,c} \neq 0, \forall i \in I_{u\tau}, c \in C\}$ is the set of items rated by u, belonging to category c;
- $C_u = \{c | C_{i,c} \neq 0, \forall c \in C, \forall i \in I_{u\tau}\}$ is the set of categories for which u has experience;
- $C_{i,c}$ denotes the occurrence of category for each item in rated above threshold τ.

4.1.3 User groups formation. We assume that users can form groups based on their similar preferences, otherwise collaborative filtering would not be possible. Moreover, by clustering the users, a basic similarity to the active user could be guaranteed, which defines also an implicit measure of recommendation relevance. Thus, by receiving the output of the category-based user profile modeling process, the steps are illustrated in Figure 5. This process, similarly to item categorization applies a variant of TFIDF used in LSA [17] to the category-based profiles, namely CFIUF (category frequency, inverse user frequency), aiming to model the similarities of users by weighting the category scores for each user. Subsequently, the normalized CFIUF-weighted user profiles are used for the clustering analysis, whose methodology is identical to the item categorization process. As a result, the groups of users are created and the recommendation procedure can be explained in detail.

4.2 Diversification Module

We divide the diversification procedure in two major steps which require the presence of the active user: *(1)* Transformation of the active user to her category-based profile; and *(2)* Formation of the distant neighborhood.

4.2.1 Active user profile modeling and classification. This module is required as the active user profile is expected to contain the raw ratings for each consumed items. Here, the role of the preprocessing modules described in the previous section is to transform the active user profile into the category-based one. Once the active user has been transformed into the CFIUF-weighted profile, the classification to a group of users is achieved through the selection of the nearest cluster, which can be computed using k-Nnearest Neighbor on the clusters.

4.2.2 Distant neighborhood formation: accuracy VS diversity. This is the pivotal step of the diversification procedure, as it involves the control of the diversity between the users within the same cluster (Figure 6) and arguably, the control of the recommendation list diversity. The most significantly diverse users within the same cluster are filtered, as opposed to nearest neighbors, according to Equation 5. As such, we expect that the categories of items from distant users will appear different but not radically, from the active user (since the prerequisite of the recommendations is to be still accurate to the active user preferences). In order to control the neighbor diversification, we introduce an external parameter α to determine the pairwise diversity significance, similarly to [15]. For a given user and another user u belonging to the same cluster c, the diversity significance s is calculated as follows:

$$s(u,v) = (1-\alpha) \cdot (1 - d(u,v)) + \alpha \cdot (d(u,v)), \quad (5)$$

where the diversity is proportional to the growth of α, hence, the larger α, the greater the diversity; $d(u,v)$ is the distance between the active user u and user v; and $(1-\alpha)$ and α respectively control the similarity and distance trade-off between the two users. Hence the k most significant users are extracted so that the significance between users u and v is maximized as:

$$\overline{V} = argmax_{v \in V}(s(u,v)) \quad (6)$$

where $\overline{V}$, set of distant neighbors, is the result of the maximized diversification significance.

5 IMPLEMENTATION AND EVALUATION

Here we address the item categorization and user profile modeling and we finish by providing the experiment setup to measure the diversification levels reached through extraction of distant neighbors. We hypothesize that the items can form more or less homogeneous categories, the users can be clustered according to such categories and finally, that the selection of distant users within the same clusters can fulfill the accuracy VS diversity trade-off.

5.1 Feature Engineering and Data Preparation

As a result of this process, the initial user and item profiles are modeled and ready to be preprocessed in the second phase according to the full recommendation procedure.

We apply the proposed diversification approach to the small Movielens dataset [18], maintained by GroupLens Research, which provides around 100K explicit ratings of 671 users for 9125 movies. Along with this dataset, we extracted the metadata from IMDb[1] to describe the items for the categorization step.

[1]The metadata have been extracted from the available interfaces at http://www.imdb.com/interfaces/.

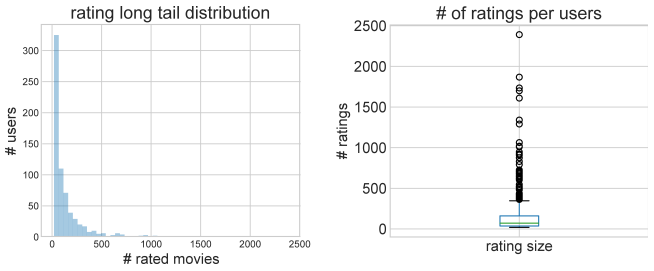

Figure 7: Movielens long tail distribution.

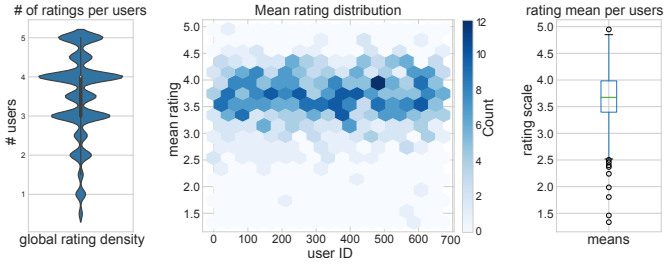

Figure 8: Movielens rating pattern distribution.

5.1.1 Characterizing user ratings in Movielens. We analyze the Movielens dataset for discovering the rating distributions, popularity biases, and differences in rating scales, to distinguish between *enthusiastic* and *demanding* users. As a result, we select the individual rating threshold to consider positive preferences. Users have rated the movies on a clear long-tail distribution, which is better depicted in Figure 7: the majority has rated less than 500 movies.

The individual rating patterns are depicted in Figure 8, through global rating density and average rating distributions. Generally, individual averages tend to be more compact near 3.5 and 4 (in rating scale), following that users tend to give high ratings, but the presence of lower and higher averages suggests and clearly proves that users have different rating behaviours. The existence of varying rating scales is also supported by [19] and reinforced by an analysis performed on the effect of the rating scale granularity [20]. Moreover, we uncover the existence of subjective scales at which users adhere when rating movies [21]: as preferences are subjective, we need to standardize ratings so that individual biases can be removed to compare rating scales objectively and finally, to extract the positive preferences. We choose the z-score standardization [21] to transform individual ratings and extract only the positive ones as we suppose that the positive preferences are simply those above the individual average. However, this approach still ignores the preferences for users having only high ratings. Therefore, we differentiate the threshold for either users with low and high standards as follows:

$$\tau_u = \begin{cases} \mu'_u & if\, \mu_u < 4 \\ 4 & if\, \mu_u \geq 4, \end{cases} \tag{7}$$

where μ_u is the average rating of u prior to standardization, so that the threshold can assume values 4 where the average is equal or above 4, and μ'_u (as the mean rating after standardization, which is always 0) for the other users.

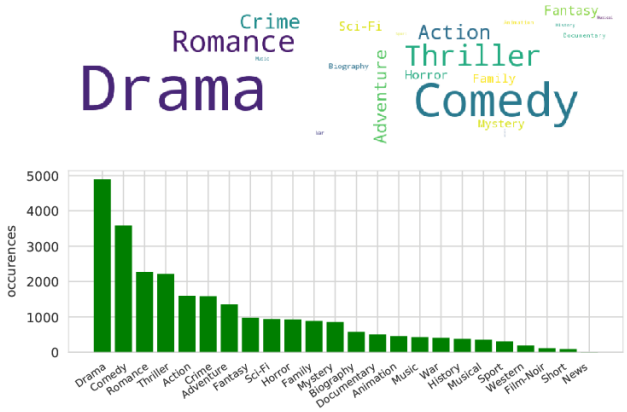

Figure 9: Genres of the movies retrieved from IMDb.

5.1.2 Characterizing movies from IMDb metadata. To provide the items with a content-based profile and not rely on ratings, we extracted descriptive metadata (*cast, company, countries, director, genres, keywords, languages, composer, release date, writer*) from IMDb, using the identifiers of the movies rated in Movielens. We notice that there is a prevalence of drama and comedy genres, which are present in around half of the extracted movies, followed by thriller and romance, present in around 20% of the movies as shown by genre inspection, Figure 9. We also stabilize the dataset, since it presented missing values after the extraction[2] by:

(1) Transforming the release date in only the year of release, from the IMDb format, complete with country, day and month of release. For movies without release date from IMDb, use the release given in the Movielens dataset.
(2) Filling missing keywords using common instances from the first occurring genre of affected movies and determine a relevance threshold.
(3) Filling missing languages and countries of production using common instances from movies with these features. Keeping the remaining missing features.

5.2 Item categorization

For this task, we sample movies with at least 3 ratings and at least an average rating of 3 from the full Movielens dataset. We then follow the process depicted in Figure 3 with the metadata extracted. The parameters utilized in the item categorization are listed in Table 1.

5.3 User group formation

Similarly to how we implemented the item categorization procedure, we model the user preferences according to Equation 4 and following Figures 4 and 5, by considering the sampled movies; thus, the number of ratings, above individual threshold account for 55416 out of 82600 for the 3685 movies considered. For this task, the parameters utilized are listed in Table 2. We select hierarchical clustering and evaluated the number of optimal clusters following

[2] We perform the stabilization process observing that a lack of features cannot be naively interpreted as an error (e.g., the lack of cast for documentary movies is due to the fact that a cast is often not required).

Table 1: Parameters and methods for item categorization.

Item Categorization parameters	values
# movies	3685
TFIDF formula	$f(t,D) \cdot \log \frac{\|D\|}{df(t)} + 1$
minimum document frequency	2
maximum document frequency	90%
k singular values (SVD)	500
distance metric	normalized Euclidean
clustering algorithm	Ward agglomerative
# clusters	43

Table 2: Parameters and methods for user groups formation.

User Groups formation parameters	values
# ratings above τ	55461
# users	671
TFIDF formula	$f(t,D) \cdot \log \frac{\|D\|}{df(t)+1}$
distance metric	normalized Euclidean
clustering algorithm	Ward agglomerative
# item categories	43
# clusters	15

the merge height elbow plot, under the criterion that the groups should allow a certain level of heterogeneity as the cluster size implicitly controls the user diversification. For the diversification approach to work, we expressively search for a clustering solution that would neither form too specific nor too generic groups of users.

5.4 Diversification Evaluation

In order to understand the diversification reached through the user clustering and distant neighborhood formation, we set up the following experimental evaluation. Specifically, our goal is to quantify how our proposed approach answers the following hypotheses:

- The approach can tune the diversification among users within the same clusters and therefore, control the trade off between user similarity and distance.
- The approach performs better than baseline diversification approaches, such as random diversification.
- The approach performs comparably to [15] as state of the art diversification method.
- The neighbors extraction with clusters produces a lower maximum yielded diversity than without (i.e. full user space).

5.4.1 Evaluation of the distant neighborhood formation. The user dataset from Movielens requires a training phase and a test phase for the offline evaluation. In the training phase, we cluster the users with the parameters listed in Table 2. Next, we classify the test user profiles on the clusters created at the training phase, by performing kNN on the closest 11 users to determine the suitable cluster. Lastly, for each of the classified test users, the distant neighborhood is formed for different values of α, with Equation 5 on the cosine distances to find the significance scores. The cosine distance is utilized as it can be easily integrated in the significance score formula, which requires both similarities and distances and more importantly, since the Euclidean distance does not have an opposite metric to calculate the similarities. We then retrieve the top$K$ neighbors according to Equation 6. Finally, we evaluate the proposed diversification approach (N) against the following methods:

- *RANDN*, the baseline method which randomly extracts the neighbors in the full user space.
- *FULLN*, extracting the neighbors in the full user space following Equation 5.
- *DNCF*, the state of the art approach from [15] which extracts distant neighbors from the active user's cluster.
- *FULLDNCF* from [15], extracting distant neighbors from the full user space.

Following the hypotheses, we expect that the user diversity can be controlled through Equation 5, considering that the evaluation at this stage allows only to measure the diversity on the distant neighbors and to reach actual recommendations, different experimental procedures are required. We expect the yielded diversity of the distant neighbors to the test users to be directly proportional to α. Moreover, since the distant neighbors are selected from the same cluster, we also expect *i)* the resulting diversity not to be as high as with random neighborhood formation and *ii)* to be less varied than with distant neighbors extracted from the full user space.

5.4.2 Evaluation procedure setup . We split the user dataset so that 80% of the users are used to train the algorithm, learn the user preference categories and form the clusters of users. The remaining 20% of users are kept for the testing phase. We use the user profiles built on item categories for the experiment. Regarding the diversification parameter α, we study the behaviour of the proposed approach for values: $\alpha = \{0, 0.2, 0.4, 0.6, 0.8, 1\}$. The values range from 0 (traditional similarity based neighbor formation, expected low heterogeneity) to varying diversification (expected low to high heterogeneity), and 1 (where the farthest users are selected, with expected high heterogeneity), respectively.

We also study the effect of the neighborhood's size with $k = \{5, 10, 15\}$ top$K$ neighbors on the yielded diversity. For the evaluation metric, we adopt the ISS metric from [15], considering the dissimilarity between pairs of users as the complement of similarity. Henceforth, the ISS metric is regarded as Intra-Set Diversity metric (ISD) and calculated as follows for the users within the neighbor set $\overline{V}_u$ of the active user u:

$$ISD_{\overline{V}_u} = 2\frac{\sum_{v,w \in \overline{V}_u, v \neq w} d(v,w)}{\left|\overline{V}_u\right| \cdot (\left|\overline{V}_u\right| - 1)} \quad (8)$$

5.4.3 Empirical results and analysis of the user diversification approaches. The results of the experiment can be inspected in Figure 10. The x-axis represents the values for which the user diversification has been conducted. On the y-axis, the resulting ISD scores are produced and represent the overall diversity of the users for the proposed approach N, together with the results of FULLN, DNCF and FULLDNCF methods, and the baseline RANDN method, which serves as the anchoring measure for the analysis. Each curve is labeled considering the significant neighbors for varying neighborhood sizes (5, 10, 15).

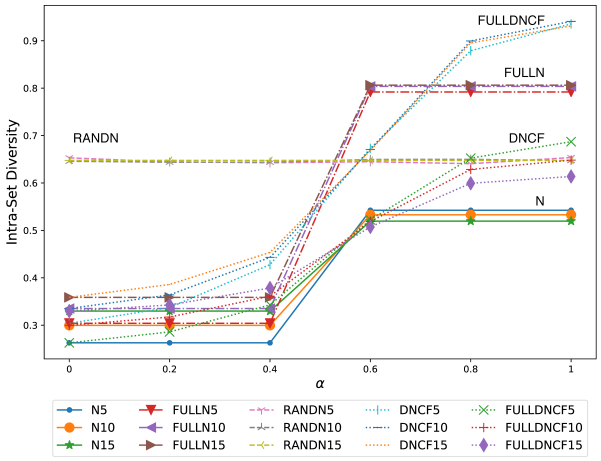

Figure 10: Intra-Set Diversity comparison of proposed approach (N) on full user space variant (FULLN), randomized approach (RANDN) and distant neighbor-based collaborative filtering (DNCF, FULLDNCF) on increasing values of α.

A general trend can be noted for the distant neighbors extracted from the same clusters (N): the proposed approach N show a crescent trend from low to high diversity, which tends to smoothen as the neighborhood size increases. Moreover, for N the ISD range varies as expected (between 0.25 and 0.52 for $k = 5$ neighbors). Instead the ISD range for the FULLN approach is wider (between 0.3 and 0.8 for $k = 5$). Surprisingly, the diversification levels do not chance as expected: instead of following a smooth curve, the changes in diversity are moderately sudden, especially for $k = 5$. In particular, there seems to be a discontinuity between $\alpha = 0.4$ and $\alpha = 0.6$ which causes this trend (for both N and FULLN), as with smaller and larger α, the ISD does not capture other variations in the neighbor diversity. This behaviour is imputable to Equation 5, which controls the neighbor extraction for the active user: with the current formula, the extracted neighbors are the same for $\alpha \leq 0.4$ and $\alpha \geq 0.6$ and therefore, the ISD scores do not vary. On the other hand, the results of RANDN, appear more constant (with ISD scores around ~ 0.64) than N and FULLN, with smaller variations in the diversity and contradictory results as the diversification level increases, producing an initially descending ISD, which increases at full diversification ($\alpha = 1$).

With the methods DNCF and FULLDNCF, the ISD scores are more consistent and varied, ranging from ~ 0.26 to 0.7 for DNCF and 0.3 to 0.9 for FULLDNCF. Compared to N, DNCF also produces the same ISD scores at $\alpha = 0$, while at maximum diversity ($\alpha = 1$), the results show a visible variation in the ISD scores: while the maximum ISD score reached by N is 0.55, it accounts at ~ 0.7 for DNCF. The FULLN and FULLDNCF methods also perform in a similar fashion (FULLDNCF has a greater maximum ISD of 0.9 compared to 0.8 of FULLN).

The results of this experiment suggest to better model the similarity VS diversity trade-off (Equation 5) to remove the evident discontinuity affecting the ISD scores. Nevertheless, we can confirm the hypothesis concerning the proportionality between the diversification levels and the diversity reached, suggesting that the proposed within-cluster neighbor diversification may be worth of consideration in the next stage of this work.

Moreover, the hypotheses on the increasing diversification levels for distant neighbors within the same cluster are valid to the extent of the maximum dissimilarity between any pair of users in the same cluster: as the users within the same groups share a theoretical baseline similarity, it would be unreasonable to expect a constant increment of the ISD scores, also considering the group sizes. To support this statement, the maximum ISD scores for the approaches (N and DNCF) using clusters of users are lower than the scores of FULLN and FULLDNCF, which operate on the full user space. As Figure 10 shows, the ISD scores are influenced by the clusters of users: the ISD for N and DNCF (within-cluster) are lower than the scores of FULLN and FULLDNCF, which operate on the full user space, conforming with our last hypothesis.

6 CONCLUSIONS AND FUTURE WORK

In this paper, we proposed a user-centric conceptual framework to control the recommendation diversity for individual users, considering relevant aspects to evaluate against individual preferences. We therefore defined diversity on the basis of four properties: local coverage (familiarity with current preferences), global coverage (the system's abilities to cover the item catalogue), novelty (unfamiliarity with current preferences) and redundancy (the amount of similar items).

Following the proposed framework, we subsequently developed a diversification procedure which can be incorporated by design into the recommendation filtering prior to the extraction of items. We developed a LSA-based user modeling based on categories of favourite items (see Figure 4). Then, we clustered groups of similar users to allow a baseline accuracy among the preferences of each group member. We evaluated our approach on the diversification of neighbors for active users by adopting the ISD metric [15] and proved that by adjusting the diversity levels of the held-out users it is possible to extract different neighbors (for both the proposed and state of the art approach) from which we can obtain a list of recommendations. Yet, the limitations of Equation 5, notably the adoption of a simple similarity VS diversity trade-off are the major cause of extracting always the same users for diversification levels $\alpha \leq 0.4$ and $\alpha \geq 0.6$, even if the ISD scores vary similarly to the state of the art method (Figure 10).

As we tested only the diversification of the distant neighborhoods, we will pursue the complete recommendation procedure and test how our approach impacts the user satisfaction in terms of recommendation quality. Also, the limitations on the preference modeling and the neighbor significance formulas suggest an area for future optimization. Moreover, we will devise a better item categorization procedure to include or remove specific metadata and base it on individual or aggregate metadata. Lastly, we will extend our framework with contextual factors (temporal changes of user preferences, time and location), as the motivations behind variety seeking behaviors comprise both internal and external factors [22]. For this purpose, we will test our framework and approach on other datasets than Movielens in conjunction with online experiments.

REFERENCES

[1] T. T. Nguyen, P.-M. Hui, F. M. Harper, L. Terveen, and J. A. Konstan, "Exploring the filter bubble," in *Proceedings of the 23rd international conference on World wide web - WWW '14*. New York, New York, USA: ACM Press, 2014, pp. 677–686. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2566486.2568012

[2] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl, "Evaluating collaborative filtering recommender systems," *ACM Transactions on Information Systems*, vol. 22, no. 1, pp. 5–53, 2004. [Online]. Available: http://portal.acm.org/citation.cfm?doid=963770.963772

[3] S. M. McNee, J. Riedl, and J. A. Konstan, "Being accurate is not enough," in *CHI '06 extended abstracts on Human factors in computing systems - CHI EA '06*, 2006, p. 1097. [Online]. Available: http://dl.acm.org/citation.cfm?doid=1125451.1125659

[4] K. Nehring and C. Puppe, "A Theory of Diversity," *Econometrica*, vol. 70, no. 3, pp. 1155–1198, 5 2002. [Online]. Available: http://doi.wiley.com/10.1111/1468-0262.00321

[5] K. JUNGE, "Diversity of ideas about diversity measurement," *Scandinavian Journal of Psychology*, vol. 35, no. 1, pp. 16–26, 3 1994. [Online]. Available: http://doi.wiley.com/10.1111/j.1467-9450.1994.tb00929.x

[6] A. Stirling, "A general framework for analysing diversity in science, technology and society," *Journal of The Royal Society Interface*, vol. 4, no. 15, 2007. [Online]. Available: http://rsif.royalsocietypublishing.org/content/4/15/707

[7] S. Vargas, L. Baltrunas, A. Karatzoglou, and P. Castells, "Coverage, redundancy and size-awareness in genre diversity for recommender systems," in *Proceedings of the 8th ACM Conference on Recommender systems - RecSys '14*. New York, New York, USA: ACM Press, 2014, pp. 209–216. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2645710.2645743

[8] P. Castells, N. J. Hurley, and S. Vargas, "Novelty and Diversity in Recommender Systems," in *Recommender Systems Handbook*. Boston, MA: Springer US, 2015, pp. 881–918. [Online]. Available: http://link.springer.com/10.1007/978-1-4899-7637-6_26

[9] C.-N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen, "Improving recommendation lists through topic diversification," in *Proceedings of the 14th international conference on World Wide Web - WWW '05*, no. January. New York, New York, USA: ACM Press, 2005, p. 22. [Online]. Available: http://portal.acm.org/citation.cfm?id=1060754%5Cnhttp://portal.acm.org/citation.cfm?doid=1060745.1060754http://portal.acm.org/citation.cfm?doid=1060745.1060754

[10] T. Murakami, K. Mori, and R. Orihara, "Metrics for Evaluating the Serendipity of Recommendation Lists," in *New Frontiers in Artificial Intelligence*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 40–46. [Online]. Available: http://link.springer.com/10.1007/978-3-540-78197-4_5

[11] M. Kaminskas and D. Bridge, "Diversity, Serendipity, Novelty, and Coverage," *ACM Transactions on Interactive Intelligent Systems*, vol. 7, no. 1, pp. 1–42, 12 2016. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3028254.2926720

[12] G. Adomavicius and Y. O. Kwon, "Toward more diverse recommendations: Item re-ranking methods for recommender systems," 2009. [Online]. Available: https://experts.umn.edu/en/publications/toward-more-diverse-recommendations-item-re-ranking-methods-for-r

[13] M. Kunaver and T. Požrl, "Diversity in recommender systems – A survey," *Knowledge-Based Systems*, vol. 123, pp. 154–162, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0950705117300680

[14] T. Aytekin and M. O. Karakaya, "Clustering-based diversity improvement in top-N recommendation," *Journal of Intelligent Information Systems*, vol. 42, no. 1, pp. 1–18, 2 2014. [Online]. Available: http://link.springer.com/10.1007/s10844-013-0252-9

[15] C. Yang, C. C. Ai, and R. F. Li, "Neighbor Diversification-Based Collaborative Filtering for Improving Recommendation Lists," in *2013 IEEE 10th International Conference on High Performance Computing and Communications & 2013 IEEE International Conference on Embedded and Ubiquitous Computing*. IEEE, 11 2013, pp. 1658–1664. [Online]. Available: http://ieeexplore.ieee.org/document/6832116/

[16] A. Carrillo, Barraza-Urbina, B. Heitmann, C. Hayes, and A. Ramos, "XploDiv: Diversification Approach for Recommender Systems," 1 2015. [Online]. Available: https://aran.library.nuigalway.ie/handle/10379/5081

[17] S. Deerwester, S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE*, vol. 41, no. 6, pp. 391–407, 1990. [Online]. Available: http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.108.8490

[18] F. M. Harper and J. A. Konstan, "The MovieLens Datasets: History and Context," *ACM Trans. Interact. Intell. Syst.*, vol. 5, no. 4, pp. 19:1–19:19, 2015. [Online]. Available: http://doi.acm.org/10.1145/2827872

[19] J. B. Schafer, D. Frankowski, J. Herlocker, and S. Sen, "Collaborative Filtering Recommender Systems," in *The Adaptive Web*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 291–324. [Online]. Available: http://link.springer.com/10.1007/978-3-540-72079-9_9

[20] F. Cena, C. Gena, P. Grillo, T. Kuflik, F. Vernero, and A. J. Wecker, "How scales influence user rating behaviour in recommender systems," *Behaviour & Information Technology*, pp. 1–20, 5 2017. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/0144929X.2017.1322145

[21] T. Hofmann and Thomas, "Latent semantic models for collaborative filtering," *ACM Transactions on Information Systems*, vol. 22, no. 1, pp. 89–115, 1 2004. [Online]. Available: http://portal.acm.org/citation.cfm?doid=963770.963774

[22] L. McAlister and E. Pessemier, "Variety Seeking Behavior: An Interdisciplinary Review," *Journal of Consumer Research*, vol. 9, pp. 311–322. [Online]. Available: https://www.jstor.org/stable/2488626