Journal of Intelligent Information Systems (2020) 54:297–316
https://doi.org/10.1007/s10844-019-00548-x

Evaluating content novelty in recommender systems

Marcelo Mendoza[1] · Nicolás Torres[1]

Received: 12 April 2018 / Revised: 4 February 2019 / Accepted: 4 February 2019 /
Published online: 9 March 2019

Abstract
Recommender systems are frequently evaluated using performance indexes based on variants and extensions of precision-like measures. As these measures are biased toward popular items, a list of recommendations simply must include a few popular items to perform well. To address the popularity bias challenge, new approaches for novelty and diversity evaluation have been proposed. On the one hand, novelty-based approaches model the quality of being new as apposed to that which is already known. Novelty approaches are commonly based on item views or user rates. On the other hand, diversity approaches model the quality of an item that is composed of different content elements. Diversity measures are commonly rooted in content-based features that characterize the diversity of the content of an item in terms of the presence/absence of a number of predefined nuggets of information. As item contents are also biased to popular contents (e.g., drama in movies or pop in music), diversity-based measures are also popularity biased. To alleviate the effect of popularity bias on diversity measures, we used an evaluation approach based on the degree of novelty of the elements that make up each item. We named this approach content novelty, as it mixes content and diversity approaches in a single and coherent evaluation framework. Experimental results show that our proposal is feasible and useful. Our findings demonstrate that the proposed measures yield consistent and interpretable results, producing insights that reduce the impact of popularity bias in the evaluation of recommender systems.

Keywords Recommender Systems · Evaluation · Diversity · Novelty

1 Introduction

Recommender systems are important tools for the discovery of unseen items on many e-commerce sites, such as Amazon, Netflix, or Spotify. The recommended items grasp a variety of different products, such as movies (Wang et al. 2014), books (Alharthi et al. 2017),

✉ Marcelo Mendoza
marcelo.mendoza@usm.cl

Nicolás Torres
nicolas.torresr@usm.cl

[1] Universidad Técnica Federico Santa María, Santiago, Chile

and songs (Chen and Chen 2005). With millions of users around the world, the development of new services for content consumption is one of the most successful business areas on the Internet. As recommender systems are key building blocks of these services to help users find novel content, they can aid in increasing service success.

Recommender systems elaborate recommendations using two kinds of methods: memory-based and model-based methods (Breese et al. 1998). Memory-based methods identify similar users or similar items to produce recommendations. User-based recommendations (Resnick et al. 1994) identify a group of similar users to a target user, retrieving the whole set of recommendations of the group and recommending advisable items from the reviewed items of these users. Analogously, item-based recommendations (Sarwar et al. 2001) identify a group of similar items to a target item, retrieving their reviews to come up with new recommendations. Model-based methods work over the whole dataset (the rating matrix of users across items) and fit a model to the data to predict the value of unobserved user-item pairs. Typical methods for model fitting employed for this purpose are probabilistic latent semantic analysis (pLSA) (Hofmann 2004), non-negative matrix factorization (MF) (Koren et al. 2009) and singular value decomposition (SVD) (Takács et al. 2007).

Another valuable approach in the design of recommender systems is the use of the content of the items to give recommendations (Lops et al. 2011). Typical approaches in this direction emanate from the detection of similar items to a given target using proximity functions based on content descriptors (Pazzani and Billsus 2007). These kinds of algorithms are known as content-based algorithms. Their main ability is to detect recommendable items when the system does not have feedback from users, favoring the recommendation of new unrated items. However, these algorithms lead to the creation of very specific user profiles, discouraging the detection of out-of-the-box items, a problem known as *over-specialization*. This issue has been recognized as a key challenge in recommender systems (Abbassi et al. 2009)

Hybrid approaches combine collaborative and content-based methods (Sarnè 2015). Hybrid approaches are based on the combination of different recommendations with voting schemes (Pazzani 1999), including collaborative filtering features in content-based approaches (Soboroff and Nicholas 2000), or including content-based features to develop recommendations with collaborative filtering (Popescul et al. 2001). A key aspect in the design of recommender systems is the ability to properly evaluate the quality of their recommendations. User feedback is utilized as the main evidence of item relevance for this purpose (Gomez and Hunt 2015). This type of evaluation is known as a single-user recommender system evaluation. When the evaluation considers groups of users, we need to segment the profiles according to specific features. Then, the performance of the recommender system is assessed in each group. This type of evaluation is known as group recommender system evaluation (Trattner et al. 2018). This article is focused on single-user evaluations.

Two information sources are commonly used in single-user recommender systems. *Implicit feedback* can be employed as proof of relevance because it indicates the level of interaction of the user with a given item. Common sources of implicit feedback are clicks or dwell time (Mendoza and Baeza-Yates 2008; Baeza-Yates et al. 2005). Click-through data can be used as a source of feedback, and the item is considered as relevant to a user if the item was selected (bought, reproduced, or viewed); otherwise, it is regarded as irrelevant. Dwell time can be employed as a graded source of feedback. The greater the time spent using the item (e.g., watching a movie, listening to a song), the more evidence there is about the relevance of the item to a given user. *Explicit feedback* is provided by users to recommender systems when they give reviews. It is common to provide a graded relevance scale to rate items. Then, recommender systems collect these ratings to improve their services. Through using either implicit or explicit feedback, it is possible to create a rating matrix of users across items where entries of the matrix are binary or graded relevance ratings. Note that the rating matrix is sparse, as it will have many unobserved user-item pairs that correspond to the unrated items of each user. The task of a recommender system is to provide a reliable rate estimation for these unobserved pairs.

Two main approaches can be employed to evaluate the quality of rating estimations. *Rating prediction* measures the precision of the estimation of the rating value. Thereafter, the evaluation is measured at the *user-item* level in terms of the error between the estimate and actual rating. Frequent measures for the assessment of rating prediction are the aggregation of absolute or squared errors. *Ranking prediction* measures the degree of accuracy of the order of the items in a list of recommendations. This approach, widely employed in information retrieval, can model more phenomena than those considered in rating prediction; among them browsing models. In this scenario, the evaluation is conducted at *top-N ranked lists*. Common measures used for this purpose are precision and recall.

Either rating or ranking prediction approaches are variants of precision-based measures, which are sensitive to the presence of popular items. As popular interests are unbalanced in terms of preferences, it is typical that a few items funnel user preferences producing a long-tail effect (Zhao et al. 2013). The long-tail effect is explained by the presence of a least-effort behavior on preferences. User preferences are ruled by the Zipf law (Dupret et al. 2006); a few items record the majority of the preferences, and a long-tail of unseen/undiscovered items appears. As user preferences are employed as a source of relevance feedback in precision-based evaluation, the existence of popular items in recommendations raises the performance of any recommender method in terms of rating or ranking prediction.

The limitation of precision-like evaluation measures is one of the key challenges in recommender systems evaluation (Bellogin et al. 2011). However, precision-like measures are still dominant in this field. This is because precision-like measures are based on relevance criteria —many times binary relevance criteria. This type of data is easy to detect (for instance, via explicit or implicit feedback) and interpret. In addition, precision-like measures are easy to compute. Saracevic (1995) stated that the challenge for evaluation is broadening of approaches and ”getting out of the isolation and blind spots of single-level, narrow evaluations”. Saracevic identified this limitation with respect to how difficult it is to integrate output levels (evaluation from the machine-side viewpoint) and user-level questions (how the user interacts with model outputs). We propose new measures to highlight different aspects of a recommender system, creating an evaluation of system outputs, going forward in content-based evaluations and bridging the gap between diversity and novelty.

Novelty and diversity are complementary concepts (Castells et al. 2015). Novelty is the quality of being new or different from what is already known. Thus, novelty is a perceived quality of an item from a user perspective. Accordingly, it is common to use views or rates as proxies for novelty estimation. Diversity is a quality of a list of items that are composed of different content elements. Thus, diversity is a perceived quality of an item from a content viewpoint.

It is typical to employ content-based features to characterize the content of an item in terms of the presence/absence of a number of predefined nuggets of information. Nuggets grasp a variety of content dimensions among them genres of music or movies.

We propose a new strategy for recommender systems evaluation by posing the concept “content novelty”. We define content novelty as the level of novelty of a list of items with respect to the contents. Content novelty combines the best of novelty and diversity approaches in a single and coherent evaluation framework, bridging the gap between both. Our goal is to measure the degree of novelty of the contents of the items in a list to infer the gain in terms of content novelty. We intend to show that content novelty approaches point to different aspects of a recommender system that diversify recommendations across users and affect the ability of the recommender system to personalize those recommendations.

The specific subject of recommender systems evaluation regarding novelty and diversity has garnered interest in the past several years. Recently, in a survey of this subject (Kunaver and Požrl 2017), the relevance of diversity for recommender systems was underscored as a key challenge. In that survey, the authors claimed that novelty and diversity are fundamental aspects of recommendation effectiveness. We believe that our proposal is relevant and timely for this subject. In particular, the introduction of our new concept, *content novelty*, will assist the community in advancing the discussion of this subject.

Recommender systems tend to overspecialize their recommendations by exploiting what we already know about the user. These kinds of recommendations can be predictable and to some extent useless. The notion of novelty and diversity indicates identifying unexpected items, giving more value to out-of-the-box recommendations. Our concern is determining how to measure unexpectedness from a wide variety of methods that over exploit similarity notions. This is a major point for recommender systems design. Content novelty will help uncover this element during the evaluation stage.

The contribution of our work can be outlined as follows. We provide new evaluation measures that can highlight differences between popularity biased methods and methods based on diversification and/or personalization. Our results will show that a new kind of measure is needed to provide a fair evaluation in terms of novelty and diversity. We define these performance measures here.

This article is organized as follows. In Section 2 we discuss related work. Our proposal is introduced in Section 3. We present and discuss our experiments in Section 4 and give our conclusions in Section 5.

2 Preliminaries

2.1 Popularity bias

Popularity bias describes the phenomenon by which a number of popular items enhance the performance of a given recommender system. Popularity bias is a well-known phenomenon in recommender systems that limits the novelty and, therefore, the quality of recommendations (Bellogin et al. 2011). In fact, many researchers use item popularity as a proxy for the inverse of novelty (Channamsetty and Ekstrand 2017).

Popularity bias arises when user ratings are concentrated in a few popular items. As user ratings are employed as a relevance feedback source, the recommendations produced using ratings are biased on popularity. We illustrate this in Movie Lens 100K (ML100K), exploring how user ratings bias user-KNN recommendations, a standard collaborative filtering method. ML100K recorded 100000 ratings by 943 users over a collection of 1664 movies. Table 1 lists some of the top movies in terms of recommendations and their presence provided by user-KNN. The full list of movies and the number of times each one was included in top-5 lists by user-KNN is shown in Fig. 1.

As Table 1 and Fig. 1 portray, popular movies are included in many lists. This is addressed by novelty measures, as novelty is expected to be optimized when novel items

Table 1 Top movies and their presence in lists produced by user-KNN

Movie	# lists	coverage on test set
Star Wars (1977)	725	76%
The Godfather (1972)	373	39%
Titanic (1997)	372	39%
Fargo (1996)	357	37&

are recommended to users. In this case, a novel item corresponds to an item that is difficult to find in a system.

The study of the effect of popularity bias in user profiles has attracted the attention of researchers in recent years. The conclusion of these studies is that many common recommender algorithms are missing a component of personalization (Channamsetty and Ekstrand 2017). Recently, Ekstrand et al. (2018) demonstrated that these limitations have effects in terms of demographic bias. However, the relation between demographic bias and popularity bias is still unclear. Algorithms to explicitly measure and respond to user profile characteristics appear as a key design feature in recommender systems. Another line of research relies on the need to focus recommender system evaluation on the diversity and novelty of the recommendations. This last challenge is the specific aspect of the problem that we address in this article.

2.2 Novelty measures

Castells et al. (2015) proposed a probabilistic framework dubbed RankSys for novelty evaluation in recommender systems. The proposed framework develops an item novelty-based approach. Basically, an item will be novel if it is difficult to find in a given dataset. The degree of difficulty in discovering a new item will depend on the ratings that an item receives. Novelty measures are calculated at the level of the list of recommendations provided to a target user u, denoted with R_u.

Browsing models are incorporated into RankSys by considering cumulative gain functions. A browser discount factor is aggregated into the gain function modeling the cost to reach a valuable item after $r-1$ recommendations. The cumulative gain function over R_u is defined by $m(R_u) = C_u \sum_{i \in R_u} \text{disc}(k_i) \cdot p(\text{rel}|i,u)$, where $p(\text{rel}|i,u)$ is the relevance rating of an item i provided by a user u, $\text{disc}(k_i)$ is a discount function and C_u is a normalization factor. Note that $m(R_u)$ allows NDCG (Järvelin and Kekäläinen 2000) to be defined

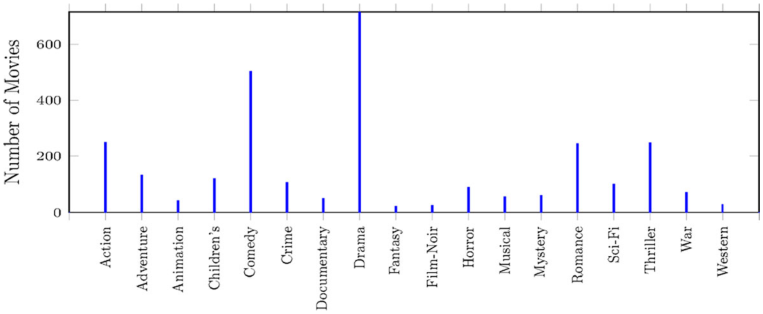

Fig. 1 Number of times each movie was recommended by `userKNN`

by replacing $\mathrm{disc}(k_i)$ with $\frac{1}{\log(r)}$. Note that $p(\mathrm{rel}|i,u) = 1$ if $(u,i) \in R_{\mathrm{test}}$. As R includes graded relevance ratings we need to include a relevance threshold ρ. The parameter ρ permits definition of the level of flexibility of the evaluation. For instance, if the relevance scale ranges in $\{1,5\}$, $\rho = 4$ indicates that $p(\mathrm{rel}|i,u) = 1$ if i was rated by u in the test set with a rating equal to or greater than 4.

Typical estimators for novelty are complements of $\frac{|U_i|}{|U|}$ or $\frac{|U_i|}{|R|}$, i.e. the fraction of users that saw i ($|U_i|$) over the total number of users of the system ($|U|$) or over the total number of ratings provided to a system ($|R|$). By combining novelty estimators with browsing models RankSys obtains a number of novelty measures. Salient measures of RankSys are "Expected Popularity Complement" (EPC) and "Expected Free Discovery" (EFD) and are defined in the following equations:

$$\mathrm{EPC}(R_u) = C_u \sum_{i \in R_u} \mathrm{disc}(k_i) \cdot p(\mathrm{rel}|i,u) \cdot \left(1 - \frac{|U_i|}{|U|}\right)$$

$$\mathrm{EFD}(R_u) = -C_u \sum_{i \in R_u} \mathrm{disc}(k_i) \cdot p(\mathrm{rel}|i,u) \cdot \log\left(\frac{|U_i|}{|R|}\right)$$

Vargas (2015) used a distance function between items to enrich RankSys extending the original definition of Intra-List Similarity (ILS) of Ziegler et al. (2005). The distance function $\mathrm{dist}(i,j)$ provided in RankSys is measured as the distance between the rating vectors of a pair of items. Next, a pair of items is close if they share ratings from the same users. This notion of distance is based on the assumption that similar items are preferred by the same users. Then, novelty estimators can be replaced with distances between items. Two salient measures based on these assumptions, "Expected Profile Distance" (EPD) and "Expected Intra-List Distance" (EILD), are defined as follows:

$$\mathrm{EPD}(R_u) = C_u \sum_{i \in R_u} \sum_{j \in I_u} \mathrm{disc}(k_i) \cdot p(\mathrm{rel}|i,u) \cdot p(\mathrm{rel}|j,u) \cdot \mathrm{dist}(i,j).$$

$$\mathrm{EILD}(R_u) = \sum_{i,j \in R_u} C_i \cdot \mathrm{disc}(k_i) \cdot \mathrm{disc}(k_j|k_i) \cdot p(\mathrm{rel}|i,u) \cdot p(\mathrm{rel}|j,u) \cdot \mathrm{dist}(i,j).$$

In the definition of EPD, the novelty of a given item i in R_u is measured from its distance to the rest of items in the profile of u, denoted by I_u. In the definition of EILD, novelty is measured in R_u, inferring the novelty of i from its distance to the rest of recommended items.

Note that all these measures are pure novelty-based measures as they are content agnostic. We will consider these measures in our experiments to illustrate differences between pure novelty approaches and our content novelty measures.

2.3 Diversity measures

As mentioned earlier, Ziegler et al. (2005) proposed ILS, a measure of the level of similarity between the items that belong to R_u. As such, ILS can be used as a proxy for the inverse of diversity. ILS can be computed from textual descriptions but suffers from the problem of the curse of dimensionality. To alleviate the effect of text sparseness, Vig et al. (2012) utilized social tags to describe movies. The approach, dubbed tag genome, infers a number of tags using a system to collect movie reviews (Movie Tuner). ILS can be computed using the tag genome (Channamsetty and Ekstrand 2017) and exhibits robust properties in terms of diversity measurement. However, movies that are not considered in the tag genome project mustbe ignored in the evaluation.

The comparison of a given list of recommended items, R_u, with an ideal list is a successful idea in evaluation that comes from the information retrieval community. Järvelin and Kekäläinen (2002) proposed the normalized discounted cumulative gain (NDCG) measure that incorporates a browsing model into the traditional precision/recall evaluation approach. The method accounted for the relative position of the documents in the ranking to discount from the gain function a factor proportional to its position in the list. Relevant documents in top positions of the list produce higher gains than those at the bottom of the list. The use of NDCG as a measure of performance in recommender systems has attracted research interest in recent times (Ekstrand et al. 2018)

Following this idea, Clarke et al. (2008) proposed a variant of NDCG for diversity evaluation, namely α-NDCG. In that approach, each item contributes to R_u with nuggets (i.e., a set of informational containers that contributes to enriching the diversity of R_u). Nuggets may correspond to music/movie/book genres or user intents in the case of web queries.

Let $\{n_1, \ldots, n_m\}$ be a set of nuggets in a dataset. Clarke et al. defined the function $N(i, n_j) = 1$ if the item i contains the n_j nugget, 0 otherwise. Then, the number of nuggets in i corresponds to $\sum_{j=1}^{m} N(i, n_j)$. To penalize the redundancy of nuggets in R_u, they defined $r_{j,K-1} = \sum_{k=1}^{K-1} N(i_k, n_j)$, which corresponds to the number of items in R_u that contains the nugget n_j, until the position $K - 1$ in the list. For convenience, $r_{j,0} = 0$. As such, the gain function at position K is defined by

$$G[K] = \sum_{j=1}^{m} N(i_K, n_j) \cdot \alpha^{r_{j,K-1}}, \quad 0 \leq \alpha \leq 1.$$

where i_K is the item in the K-th position of R_u. If we want to increase the relative weight of nugget redundancy for gain discount, we need to take small values for α. By applying a discount function, we can obtain a α-DCG curve for R_u. Finally, a normalized version α-NDCG is obtained by comparing α-DCG with an ideal list in terms of diversity. The ideal list is greedy with regard to the number of nuggets, maximizing the cumulative gain at each level of R_u.

We may note that α-NDCG is an extension of NDCG, but it incorporates content as a key element for evaluation. Thus, the point behind α-NDCG is to provide a measure that takes into account the content of the item, a disruptive approach in information retrieval where for decades the dominant measures were based on relevance criteria (for instance, binary relevance) and not explicitly on content. We highlight the contribution of α-NDCG as a pure diversity-oriented measure, in the sense that the primary guideline of the measure is content. We will build our proposal extending α-NDCG to consider novelty aspects, a necessary effort to bridge the gap between novelty and diversity measures.

3 Content novelty measures for recommender systems

We propose a number of content novelty measures for recommender systems. Our proposal seeks to bridge the gap between diversity measures (e.g. α-NDCG) and novelty measures (e.g., EPC, EFC). The concept is to apply the novelty approach in the context of diversity analysis, an approach that we call content novelty. We do this by taking advantage of content nuggets, reducing the effect of the bias that comes from user ratings. However, we will need to deal with genre-bias to produce unbiased estimators of content novelty.

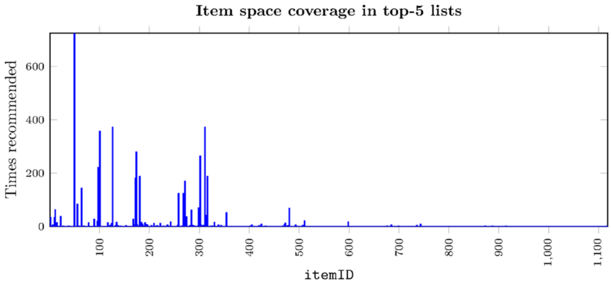

Fig. 2 Number of movies per genre in ML100K

Genre-bias We will use Movie Lens 100K (ML100K) as a running example to illustrate the effect of genre-bias on recommendations. On ML100K, movies are tagged across genres by experts, and many movies feature more than one tag, demonstrating that the content correlates with more than one genre. Movie Lens considers 19 genres. We depict in Fig. 2 the number of movies per genres in ML100K. Note that ML100K is a multi-label dataset.

As Fig. 2 shows, many movies are tagged into the three majority genres (Drama, Comedy or Action) and just a few movies are tagged into Fantasy, Film Noir or Western, the minority genres on ML100K. Using genres as content nuggets, it is expected that many of the recommended lists include common genres, such as Drama, and only a few lists include Western or Fantasy. Note that genre bias is produced by experts when they recognize a genre in an item. Then, tags are biased toward specific genres because the production of items in those genres is high. Determining how to consider this bias during the evaluation process is a crucial issue for our proposal.

The multi-label nature of the process is illustrated in Table 2, where the distribution of number of genres in ML100K is shown.

When user-based collaborative filtering (user-KNN [1]) was applied to ML100K, most of the top-5 recommendation lists were concentrated into two genres, as seen in Table 3.

Table 3 shows that most of the lists include Drama or Action movies into its recommendations. Conversely, only five lists include Documentaries and only four include Fantasy movies, evidencing the presence of genre bias on user-KNN recommendations. Surprisingly, when we applied α-NDCG using genres as content nuggets, we achieved a measurement of 0.95 on top-5 lists, an almost perfect achievement. As α-NDCG is defined in terms of the number of nuggets in the lists of recommendations, it yields high values even when certain genres are underrepresented in the lists. We will show that the inclusion of novelty measures during nugget occurrences will alleviate the effect of genre-bias for evaluation purposes. A recommendation will be useful in terms of content novelty depending on the degree of difficulty in discovering a new nugget in a given system.

Genre bias and popularity bias are related. Popular movies are concentrated into a few genres (Drama and Action, or both). Therefore, user ratings reinforce the effect of genre-bias on recommendations. As user ratings are concentrated in a few movies and these movies belong to several genres, a measure like α-NDCG is biased by the combined effect of both factors. By measuring novelty at the nugget level, we separate both factors in the evaluation.

[1] User-KNN was implemented with 50 nearest neighbors using cosine similarity as proximity function.

Table 2 Distribution of genres in ML100K. Many movies are tagged into two or more genres by experts

# genres	1	2	3	4	5	6
# movies	824	563	212	51	11	3

Now, it is necessary that we discuss some concerns about popularity and genre bias. More popularity does not necessarily mean better quality. A number of factors can explain why popularity exists, and quality is just one among many factors that can help to explain the popularity phenomena. Among the relevant factors that aid in describing movie popularity, we highlight promotion, presence/absence of distracting factors (competitive events during critical opening weekends), competition during the film's release period, and popularity of the film's stars, among others (Elberse 2007). Along these lines, another point arises and it is even more crucial. If a recommender system does not take into account genre and popularity bias, these sorts of recommendations will be limited to following trends. We will show that content novelty evaluation is crucial to uncovering these aspects in recommender systems.

Personalization A first performance measure that we will discuss intends to measure personalization across lists. We will model this as a factor of α-NDCG, rewarding lists with items recommended to only one person. Let $V(u_i)$ be a function that is equal to 1 if at least one item recommended to u_i was **only** recommended to u_i, 0 otherwise. We define total diversity (Tot Div) as a global measure of diversity in a testing set, calculated as the product of α-NDCG and the fraction of lists that have at least one unique recommended item:

$$\text{Tot Div} = \alpha\text{-NDCG} \cdot \frac{\sum_{i=1}^{n} V(u_i)}{n}, \tag{1}$$

where n is the number of lists included in the testing set. Total diversity penalizes α-NDCG when the lists tend to include only popular items. Tot Div will be equal to α-NDCG if and only if all the lists in the testing set includes at least one unique item. Otherwise, Tot Div will be a fraction of α-NDCG, which corresponds to the proportion of lists with unique items over the total number of lists in the testing set. Tot Div is a measure able to detect if a recommender system produces diversification across users. However, note that the cross effect between users is not evaluated, just the effort of a recommender system to spread different items across users. The degree of success that a recommender system has in spreading items with meaningful recommendations to users is a novel aspect in recommender systems evaluation.

Content novelty on *I* A second performance measure we propose includes a discount for frequent nuggets regarding the whole set of items, I. The idea is to include a discount factor with α-NDCG to penalize the inclusion of frequent nuggets, a kind of nugget novelty

Table 3 Distribution of genres in lists recommended by user-KNN in ML100K. Most of the lists are concentrated into two major genres

Genres	Drama	Action	···	Documentary	Fantasy	Western
# lists	932 (97.6%)	849 (88.9%)	···	5 (0.5%)	4 (0.4%)	2 (0.2%)

approach applied to diversity measures. As a context for novelty we use I, the set of items in the system. We employ the nugget counting function $N(d_i, n_j)$ used in α-NDCG to define this factor. The reciprocal of a nuggets count for a given nugget, n_j, is $\frac{N}{\sum_{i=1}^{N} N(d_i, n_j)}$, where N is the number of items in I. An infrequent nugget, for instance a nugget contained in only one item, achieves a maximum value in this factor (N). On the other hand, the minimum value for this factor is achieved when the nugget is included in all the items of I (1). We use the logarithm of this factor, which resembles the IDF factor used in information retrieval. Then, we define an exponential discount factor, $\exp_1^j$ for a given nugget n_j:

$$\exp_1^j = \log_N \left[\frac{N}{\sum_{i=1}^{N} N(d_i, n_j)} \right] - 1. \tag{2}$$

A novel nugget, n_j (e.g., a nugget included in only one item of I), reaches $\exp_1^j = 0$. Frequent nuggets n_j will reach negative values, with a minimum achieved for a nugget, n_j, included in all the items of I, that corresponds to $\exp_1^j = -1$. As $\exp_1^j$ ranges in $[-1, 0]$, we use it as an exponential discount factor for $\beta \geq 1$, defining the $\alpha\beta$-NDCG diversity measure as follows:

$$\alpha\beta\text{-NDCG} = \sum_{j=1}^{m} N(d_k, n_j) \cdot \alpha^{r_j, k-1} \cdot \beta^{\ \exp_1^j}, \qquad 0 \leq \alpha \leq 1, 1 \leq \beta. \tag{3}$$

As for novel nuggets, $\exp_1^j = 0$, $\alpha\beta$-NDCG = α-NDCG. However, as for frequent nuggets, $\exp_1^j < 0$, $\alpha\beta$-NDCG = α-NDCG $\cdot \frac{1}{\beta^{|e_1^j|}}$, a fraction of α-NDCG defined by the inverse frequency of the nugget on I.

Content novelty in I_u A third performance measure we propose includes a discount for frequent nuggets on I_u (the set of items seen/rated by u). The concept behind this measure is to change the context for nugget novelty to the user profile, I_u (items seen/rated by u), measuring nugget novelty at the user level. In this way, frequent nuggets on I_u will be penalized by a discount factor, and infrequent nuggets on I_u will produce rewards in terms of α-NDCG. Infrequent nuggets on I_u represent novel contents for the user (e.g., novel genres for u). A highly recommended item with infrequent nuggets represents an unexpected recommended item in terms of user profile. To measure this effect, we define an exponential discount factor, $\exp_2^j$ for a given nugget n_j:

$$\exp_2^j = \log_{1+N_u} \left[\frac{1 + N_u}{1 + \sum_{i=1}^{N_u} N(d_i, n_j)} \right] - 1. \tag{4}$$

where N_u is the number of items seen/rated by u (items in I_u). As in I_u, $N(d_i, n_j)$ could be zero, so we shifted the function by one unit to avoid a division by zero in the logarithm. Note that the documents considered in $\exp_2^j$ ranges I_u, a significant difference with $\exp_1^j$, where the documents range I. Then, we define $\alpha\gamma$-NDCG:

$$\alpha\gamma\text{-NDCG} = \sum_{j=1}^{m} N(d_k, n_j) \cdot \alpha^{r_j, k-1} \cdot \gamma^{\ \exp_2^j}, \qquad 0 \leq \alpha \leq 1, 1 \leq \gamma. \tag{5}$$

Novel nuggets in I_u will produce high values in $\alpha\gamma$-NDCG, reaching a maximum value when the list of recommended items includes only novel nuggets for u. In this case,

$\alpha\gamma$-NDCG = α-NDCG. Yet, frequent nuggets will introduce discounts on this measure, and $\alpha\gamma$-NDCG will only reach a fraction of α-NDCG defined by the inverse frequency of the frequent nuggets on I_u.

Finally, a fourth measure that can be used for evaluation contains a combination of the previous measures. We consider $\alpha\beta\gamma$-NDCG which combines content novelty at the I and I_u levels in a single measure:

$$\alpha\beta\gamma\text{-NDCG} = \sum_{j=1}^{m} N(d_k, n_j) \cdot \alpha^{r_j,k-1} \cdot \beta^{\ \mathrm{exp}_1^j} \cdot \gamma^{\ \mathrm{exp}_2^j}, \qquad 0 \leq \alpha \leq 1, 1 \leq \beta, 1 \leq \gamma \quad (6)$$

and $\alpha\beta\gamma$ - Tot Div, which combines personalization and content novelty at I and I_u:

$$\alpha\beta\gamma - \textsc{Tot Div} = \sum_{j=1}^{m} N(d_k, n_j) \cdot \alpha^{r_j,k-1} \cdot \beta^{\ \mathrm{exp}_1^j} \cdot \gamma^{\ \mathrm{exp}_2^j} \cdot \frac{\sum_{i=1}^{n} V(u_i)}{n}, \quad (7)$$

with $0 \leq \alpha \leq 1, 1 \leq \beta, 1 \leq \gamma$.

Note that $\alpha\beta$-NDCG (3), $\alpha\gamma$-NDCG (5), and $\alpha\beta\gamma$-NDCG (6) are single-user performance measures. By averaging these measures over users we achieve a global performance measure for the recommender system. Note that Tot-Div and $\alpha\beta\gamma$ - Tot Div are global performance measures by definition, as Tot-Div is defined at the level of the whole system.

4 Examples and experiments

We start this section showing illustrative examples of how our performance measures work. We then present our test of the proposed measures, employing three different datasets with six different recommendation methods.

4.1 Illustrative examples

Let us consider the following set of movies with their respective genres:

v_1: The Silence of the Lambs(1991)::Crime|Horror|Thriller
v_2: Titanic(1997)::Drama|Romance
v_3: It's a Wonderful Life(1946)::Drama|Fantasy|Romance
v_4: Unforgiven(1992)::Drama|Western
v_5: Pulp Fiction(1994)::Comedy|Crime|Drama
v_6: The Godfather(1972)::Crime|Drama
v_7: Forrest Gump(1994)::Comedy|Drama|Romance|War
v_8: Goodfellas(1990)::Crime|Drama
v_9: The Shawshank Redemption(1994)::Drama
v_{10}: Schindler's List(1993)::Drama|War

High content novelty, low total diversity Let us assume that we have three top-1 lists (L_1, L_2 and L_3) that show high novelty in terms of genres, including Crime, Horror and Thriller movies (none of these genres have been seen by these users) and low total diversity (the three lists recommend the same movie).

User	Recommendations	I_u
u_1	L_1: $\{v_1\}$, $\langle$Crime, Horror, Thriller$\rangle$	$\{v_2, v_3\}$
u_2	L_2: $\{v_1\}$, $\langle$Crime, Horror, Thriller$\rangle$	$\{v_4, v_7\}$
u_3	L_3: $\{v_1\}$, $\langle$Crime, Horror, Thriller$\rangle$	$\{v_9, v_{10}\}$

Note that the three genres tagged in v_1 are new for users u_1, u_2 and u_3 (high novelty recommendations in terms of genres). As the recommendation is novel, the variants of α-NDCG achieve high values in spite of the fact that the lists have low total diversity (the three lists recommend the same movie). Specifically, $\alpha\gamma$-NDCG reaches the maximum value as these genres are novel in I_u. However, the low total diversity is penalized by Tot Div and $\alpha\beta\gamma$ − Tot Div. Content novelty measures (averaged) for this case are shown below.

α-NDCG	$\alpha\beta$-NDCG	$\alpha\gamma$-NDCG	$\alpha\beta\gamma$-NDCG	Tot Div	$\alpha\beta\gamma$ − Tot Div
1	0.886	1	0.886	0	0

Low content novelty, high total diversity Let us assume that we have three top-1 lists (L_1, L_2 and L_3) that show low content novelty (the three users have seen a drama movie) and high total diversity (the three lists include a different movie).

User	Recommendations	I_u
u_1	L_1: $\{v_9\}$, $\langle$Drama$\rangle$	$\{v_1, v_2\}$
u_2	L_2: $\{v_6\}$, $\langle$Crime, Drama$\rangle$	$\{v_1, v_2\}$
u_3	L_3: $\{v_8\}$, $\langle$Crime, Drama$\rangle$	$\{v_1, v_2\}$

Content novelty measures (averaged and at the user level) for this case are shown below.

	α-NDCG	$\alpha\beta$-NDCG	$\alpha\gamma$-NDCG	$\alpha\beta\gamma$-NDCG	Tot Div	$\alpha\beta\gamma$-Tot Div
u_1	1	0.516	0.649	0.335	–	–
u_2	1	0.587	0.649	0.380	–	–
u_3	1	0.587	0.649	0.380	–	–
avg	1	0.564	0.649	0.365	1	0.365

As the recommendation is diverse in terms of personalization, Tot Div achieves its maximum value. However, the content novelty measures reach low values because in terms of genres, the three lists are not novel and include Drama in three lists and Crime in two. Note that $\alpha\beta$-NDCG reaches lower values than $\alpha\gamma$-NDCG as Crime is less frequent in I than in I_u.

Recommending to experts (low novelty) Now suppose that we have three users with many movies seen/rated. Recommendations in this situation will achieve low content novelty. Let us assume that the three lists (L_1, L_2 and L_3) have high total diversity (the three lists include a different movie).

User	Recommendations	I_u
u_1	L_1: $\{v_9\}$, $\langle$Drama$\rangle$	$\{v_1, v_2, v_3, v_4, v_5, v_6, v_7, v_8, v_{10}\}$
u_2	L_2: $\{v_6\}$, $\langle$Crime, Drama$\rangle$	$\{v_1, v_2, v_3, v_4, v_5, v_7, v_8, v_9, v_{10}\}$
u_3	L_3: $\{v_8\}$, $\langle$Crime, Drama$\rangle$	$\{v_1, v_2, v_3, v_4, v_5, v_6, v_7, v_9.v_{10}\}$

Table 4 Statistics of the data sets

	ML100K	ML1M	MovieTweetings
Users (N)	943	6040	49,572
Items (M)	1664	3706	28,377
Ratings (nnzs of $\mathcal{R}$)	99,392	1,000,208	618,051
min $\|\mathcal{R}_u\|^a$	20	20	15
Density^b	6.33%	4.47%	0.04%

[a] Minimum number of ratings per user

[b] Non-zero entries (nnzs of R) over $M \times N$

Note that the three lists are different and each user has seen/rated all the movies in I except the recommended movie. Content novelty measures (averaged and at the user level) for this case are shown below.

	α-NDCG	$\alpha\beta$-NDCG	$\alpha\gamma$-NDCG	$\alpha\beta\gamma$-NDCG	Tot Div	$\alpha\beta\gamma$-Tot Div
u_1	1	0.516	0.519	0.268	–	–
u_2	1	0.587	0.613	0.367	–	–
u_3	1	0.587	0.613	0.367	–	–
avg	1	0.564	0.582	0.334	1	0.334

As the three lists are different, Tot Div reaches its maximum value. However, as the recommended movie is not novel for the users, content novelty measures penalize this situation and result in low values. The lowest value is achieved by $\alpha\beta\gamma$-NDCG showing that this measure is the more strict measure in terms of content novelty.

In summary, low novelty recommendations will be penalized by $\alpha\beta$-NDCG and $\alpha\gamma$-NDCG measures while low total diversity measures will be penalized by Tot Div. Combined measures as $\alpha\beta\gamma$-NDCG or $\alpha\beta\gamma$-Tot Div will evaluate all these factors in a single measure.

4.2 Experimental results

In this section, we compare the results obtained by the proposed measures in three different datasets. In Table 4, we show statistics for each dataset used in our experiments.

Table 4 shows that MovieTweetings is the largest dataset considered in our experiments, with the lowest density and the highest number of users and items. Conversely, ML100K is the smallest dataset with the highest density and the least number of users and items.

We will compare our measures employing six different methods of recommendations. We evaluate recommendations provided by random lists (Rnd), popularity-sorted lists (Pop), Non-Negative Matrix Factorization (MF) (Koren et al. 2009), Probabilistic Latent Semantic Analysis (PLSA) (Hofmann 2004), User-based Collaborative Filtering (UB) (Resnick et al. 1994) and Item-based Collaborative Filtering (IB) (Sarwar et al. 2001). We used RecommenderLab (Hahsler 2016) to conduct the experiments, a library of recommender systems provided in R. Results for ML100K, ML1M and MovieTweetings are found in Tables 5, 6, and 7, respectively. Bold fonts indicate best results for each measure.

Tables 5, 6, and 7 show that α-NDCG is the most permissive measure used, achieving the best results for all the methods across the three datasets. In fact, both random and popularity-based methods produce values that are very close to those accomplished with other methods.

Table 5 Results of content novelty measures for ML100K

	α-NDCG	$\alpha\beta$-NDCG	$\alpha\gamma$-NDCG	$\alpha\beta\gamma$-NDCG	TOT DIV	$\alpha\beta\gamma$-TOT DIV
Rnd	0.8221	0.6030	0.7240	0.4408	0.2461	0.1084
Pop	0.7393	0.6012	**0.7441**	**0.4576**	0.0038	0.0017
MF	0.8483	0.6098	0.7157	0.4406	**0.5962**	**0.2625**
pLSA	**0.8604**	**0.6107**	0.7077	0.4357	0.1892	0.0824
UB	0.8399	0.6106	0.7153	0.4407	0.5849	0.2577
IB	0.8314	0.6037	0.7221	0.4408	0.5607	0.2472

In particular, random lists have nearly the same performance in terms of α-NDCG as UB or IB, with only one point of difference in ML100K, and almost identical performance as IB with respect to MovieTweetings. Popularity had the worst results for α-NDCG across the three datasets. Content novelty measures are more strict. Specifically, the most strict measure of performance is $\alpha\beta\gamma$-NCDG and the most permissive measure is $\alpha\gamma$-NDCG. Meanwhile, TOT DIV is the measure that exhibits the greatest performance variance across the different methods evaluated. According to TOT DIV, the worst method is popularity, and this was as expected. Note that popularity can be seen as a proxy of the inverse of novelty. Two methods achieve the best results in this evaluation, MF and UB. While MF is better with both MovieLens datasets, UB performs better with MovieTweetings. This fact suggests that the density of the dataset affects the performance of the methods in terms of TOT DIV. According to TOT DIV, MF can diversify results in dense datasets. However, in a sparse dataset, such as MovieTweetings, UB diversifies better than MF.

When $\alpha\beta\gamma-$TOT DIV is employed, both UB and MF are penalized, indicating that strong performance in terms of TOT DIV does not imply robust performance regarding content novelty. Indeed, for this last measure, the difference between UB, MF and IB diminishes to only one point in Ml100K and ML1M. The difference in favor of MF and UB increases in MovieTweetings. In general, these results show that $\alpha\beta\gamma -$ TOT DIV allows the evaluation, with a single measure, of both aspects of a recommender system method (i.e., determines the degree of novelty of the items of a given list and the extent to which the recommended lists are different).

In Table 8, the results achieved with RankSys, the framework proposed by Vargas for novelty evaluation (Vargas 2015), are presented. In these experiments, we used the four measures discussed in Section 2. Table 8 shows that the four measures achieved differential results. EPC exhibited high novelty for recommendations produced for MovieTweetings, with almost all the methods achieving the maximum value in this measure. Surprisingly,

Table 6 Results of content novelty measures for ML1M

	α-NDCG	$\alpha\beta$-NDCG	$\alpha\gamma$-NDCG	$\alpha\beta\gamma$-NDCG	TOT DIV	$\alpha\beta\gamma$-TOT DIV
Rnd	0.8008	0.5951	**0.7181**	**0.4316**	0.0749	0.03234
Pop	0.7894	0.5779	0.7031	0.4138	0.0012	0.0005
MF	0.8343	0.5999	0.6845	0.4143	**0.3002**	**0.1244**
pLSA	**0.8577**	**0.6038**	0.6863	0.4178	0.0462	0.0193
UB	0.8365	0.6001	0.6841	0.4141	0.2996	0.1241
IB	0.8379	0.5969	0.7052	0.4254	0.2513	0.1069

Table 7 Results of content novelty measures for MovieTweetings

	α-NDCG	$\alpha\beta$-NDCG	$\alpha\gamma$-NDCG	$\alpha\beta\gamma$-NDCG	Tot Div	$\alpha\beta\gamma$-Tot Div
Rnd	0.8718	**0.5840**	**0.8110**	**0.4788**	0.2714	0.1299
Pop	0.8012	0.5742	0.7680	0.4448	0.0006	0.0002
MF	0.9047	0.5791	0.7368	0.4299	0.6125	0.2633
pLSA	**0.9169**	0.5697	0.7288	0.4250	0.1982	0.0843
UB	0.9048	0.5791	0.7368	0.4299	**0.6128**	**0.2635**
IB	0.8742	0.5812	0.7783	0.4573	0.3723	0.1702

MF achieved poor results in terms of EPC for MovieTweetings, with worse results than those obtained by random lists or even popularity. EPC produced less novelty on recommendations with ML100K and ML1M, providing the best and worst results achieved by random lists and popularity, respectively. EFD had very similar results to EPC, introducing a scaling factor that helps to increase the separation between the results obtained by the different methods. When EPD is used, the best results are achieved by popularity, indicating exactly the opposite of EPC and EFD with ML100K and ML1M. A more strict evaluation is achieved when EILD is used, accomplishing similar results as EPC and EFD but with lower performance values. For both EILD and EPD measures, the best methods are random lists and popularity-based recommendations. In particular, for MovieTweetings the best result in terms of EPD is achieved by popularity, suggesting an almost perfect performance of popularity in terms of novelty. Thus, for these datasets, the use of distance functions with EPD

Table 8 Results of the unified framework RankSys for ML100K, ML1M and MovieTweetings. Whilst EPD is biased by popularity, randomization is able to maximize EPC and EILD across almost all datasets

		EPC	EFD	EPD	EILD
ML100K	Rnd	**0.9556**	**6.8384**	0.8325	**0.7400**
	Pop	0.7660	2.5209	**0.8436**	0.6968
	MF	0.7765	2.2391	0.7634	0.6545
	pLSA	0.8633	3.1743	0.7584	0.6486
	UB	0.7999	2.4903	0.7743	0.6734
	IB	0.8382	2.9080	0.7519	0.6295
ML1M	Rnd	**0.9797**	**7.3859**	0.8305	**0.7379**
	Pop	0.7074	1.7839	**0.8415**	0.6948
	MF	0.7961	2.4184	0.7615	0.6527
	pLSA	0.8420	2.9390	0.7603	0.6486
	UB	0.7802	2.3057	0.7762	0.6753
	IB	0.9076	4.1094	0.7361	0.6061
MovieTweetings	Rnd	**0.9999**	13.0805	0.9448	0.8995
	Pop	0.9977	8.7958	**0.9883**	**0.9000**
	MF	0.9869	10.8246	0.8852	0.8653
	pLSA	0.9997	12.5692	0.8964	0.8953
	UB	**0.9999**	**13.2147**	0.9871	0.8960
	IB	0.9998	13.0070	0.9625	0.8966

Table 9 Wilcoxon signed-rank test results (p-values) for ML100K

	$\alpha\beta$-NDCG	$\alpha\gamma$-NDCG	$\alpha\beta\gamma$-NDCG	Tot Div	$\alpha\beta\gamma$-Tot Div	EPC	EFD	EPD	EILD
α-NDCG	0.03	**0.06**	0.03	0.03	0.03	**0.84**	0.03	**0.56**	0.03
$\alpha\beta$-NDCG	–	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03
$\alpha\gamma$-NDCG	–	–	0.03	0.03	0.03	0.03	0.03	0.03	**0.06**
$\alpha\beta\gamma$-NDCG	–	–	–	**0.43**	0.03	0.03	0.03	0.03	0.03
Tot Div	–	–	–	–	0.03	0.03	0.03	0.03	0.03
$\alpha\beta\gamma$-Tot Div	–	–	–	–	–	0.03	0.03	0.03	0.03
EPC	–	–	–	–	–	–	0.03	0.03	0.03
EFD	–	–	–	–	–	–	–	0.03	0.03
EPD	–	–	–	–	–	–	–	–	0.03

or EILD does not permit the detection of differences with EPC and EFD. In the following, bold fonts indicate best results for each measure.

4.3 Analysis of results

The measures analyzed in the previous section yield information about different aspects of the recommendation methods for each dataset. To verify the level of dependency between them, we conducted a Wilcoxon signed-rank test for every pair of measures in each dataset. Tables 9, 10, and 11 show the results of these tests.

Small p-values indicate that the null hypothesis is rejected. Tables 9–11 also indicate that most of the comparisons rejected the null hypothesis, and denoted with bold fonts are the cases where the evidence does not allow rejection of the null hypothesis. Accordingly, the test indicates that both results come from the same population. In Table 9, where the results for ML100K are shown, the test indicates a strong dependency between the results of α-NDCG and $\alpha\gamma$-NDCG as well as EPC and EPD. A strong dependency is also observed between $\alpha\gamma$-NDCG and EILD, along with between $\alpha\beta\gamma$ and $\textsc{Tot} - Div$. Both $\alpha\beta\gamma$ – Tot Div and EFD are the only measures that reject the null hypothesis across all comparisons, as is depicted in Table 9.

Table 10 Wilcoxon signed-rank test results (p-values) for ML1M

	$\alpha\beta$-NDCG	$\alpha\gamma$-NDCG	$\alpha\beta\gamma$-NDCG	Tot Div	$\alpha\beta\gamma$-Tot Div	EPC	EFD	EPD	EILD
α-NDCG	0.03	0.03	0.03	0.03	0.03	**1**	0.03	**0.15**	0.03
$\alpha\beta$-NDCG	–	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03
$\alpha\gamma$-NDCG	–	–	0.03	0.03	0.03	0.03	0.03	0.03	**0.15**
$\alpha\beta\gamma$-NDCG	–	–	–	0.03	0.03	0.03	0.03	0.03	0.03
Tot Div	–	–	–	–	0.03	0.03	0.03	0.03	0.03
$\alpha\beta\gamma$-Tot Div	–	–	–	–	–	0.03	0.03	0.03	0.03
EPC	–	–	–	–	–	–	0.03	**0.21**	0.03
EFD	–	–	–	–	–	–	–	0.03	0.03
EPD	–	–	–	–	–	–	–	–	0.03

Table 11 Wilcoxon signed-rank test results (p-values) for MovieTweetings

	$\alpha\beta$-NDCG	$\alpha\gamma$-NDCG	$\alpha\beta\gamma$-NDCG	Tot Div	$\alpha\beta\gamma$-Tot Div	EPC	EFD	EPD	EILD
α-NDCG	0.03	0.03	0.03	0.03	0.03	0.03	0.03	**0.15**	**0.68**
$\alpha\beta$-NDCG	–	0.03	0.03	**0.15**	0.03	0.03	0.03	0.03	0.03
$\alpha\gamma$-NDCG	–	–	0.03	0.03	0.03	0.03	0.03	0.03	0.03
$\alpha\beta\gamma$-NDCG	–	–	–	0.03	0.03	0.03	0.03	0.03	0.03
Tot Div	–	–	–	–	0.03	0.03	0.03	0.03	0.03
$\alpha\beta\gamma$-Tot Div	–	–	–	–	–	0.03	0.03	0.03	0.03
EPC	–	–	–	–	–	–	0.03	0.03	0.03
EFD	–	–	–	–	–	–	–	0.03	0.03
EPD	–	–	–	–	–	–	–	–	0.03

For ML1M, Table 10 depicts a strong dependency between α-NDCG, EPC and EPD. A strong dependence is also observed between $\alpha\gamma$-NDCG and EILD as well as between EPC and EPD. For this evaluation, our measures $\alpha\beta\gamma$-NCDG, Tot Div, and $\alpha\beta\gamma$-Tot Div rejected all the null hypotheses. From RankSys, EFD rejected the null hypothesis for all comparisons.

For MovieTweetings, almost all the compared measures yielded different results. As Table 11 shows, the only paired dependencies are seen between α-NDCG and EPD, EILD and between $\alpha\beta$-NCDG, and Tot Div. $\alpha\gamma$-NDCG, $\alpha\beta\gamma$-NDCG and $\alpha\beta\gamma$-TotDiv rejected the null hypothesis across all comparisons. From RankSys, both EPC and EFD rejected the null hypothesis for all comparisons, as well.

The results for Tables 9, 10 and 11 show that the number of dependencies between performance measures decreases when dataset complexity increases. On the one hand, for ML100K, which is the smallest dataset considered in our experiments that exhibits the highest data density, only two measures rejected the hypothesis test in all comparisons, indicating that the variability between the measures is small. Conversely, for Movie Tweetings, which is the largest dataset considered in our experiments with the lowest data density, five measures rejected the test for all comparisons, suggesting high variability. When the results across the three datasets were compared, $\alpha\beta\gamma$-Tot Div and EFD are the only ones that rejected the null hypothesis for all comparisons, showing that consistently, these measures yield results that are different than the results obtained by the other measures.

4.4 A map of methods

Figure 3 compares the overall performance of the methods in terms of our analysis (performance averaging over the datasets). In particular, portrayed is a comparison between $\alpha\beta\gamma$-TotDiv and EFD, showing that MF and UB are able to maximize $\alpha\beta\gamma$-TotDiv. While RND maximizes EFD, IB reaches the best balance between both factors. On the other hand, POP offers the poorest results when both factors are considered. These results indicate that a fair analysis needs to combine many elements. The isolation of a factor may lead to the wrong conclusions about performance.

Our results show that $\alpha\beta\gamma$-TotDiv is especially suited for measuring a different dimension of novelty —compared to the measures of Vargas (2015)— by combining nugget-aware diversity with the effect of personalization.

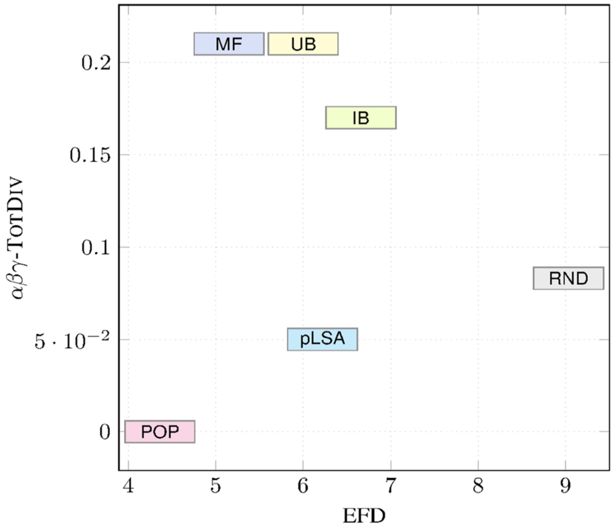

Fig. 3 Map of methods according to EFD and $\alpha\beta\gamma$-nDCG Tot Div

5 Conclusions

We proposed a set of new measures for recommender system performance evaluation. Our measures are based on a new approach for evaluation, which we dubbed content novelty. This approach measures the degree of novelty of the contents proposed in a list. To measure content novelty, we used information nuggets, an approach previously explored in information retrieval (Clarke et al. 2008). We put forth three variants of the well-known α-NDCG measure. In addition, we proposed an intra-list measure referred to as Total Diversity (Tot Div), which accounts for personalization with a set of lists. We compared our measures with α-NDCG and four measures posited in RankSys (Vargas 2015), a framework for the evaluation of novelty in recommender systems. Experiments with three different datasets employing six different recommendation methods show that our measures yield consistent and interpretable results. Considering RankSys, while EPD is biased by popularity, randomization is able to maximize EPC and EILD across almost all the datasets. When our measures were employed in the experiment, the effect of popularity bias was only observed with the smallest dataset (ML100K), while randomization was able to maximize our variants of α-nDCG in ML1M and MovieTweetings. The inclusion of Tot Div is a key factor for the success of our approach. According to Tot Div, the worst method is popularity, and this was as expected. Note that popularity can be seen as a proxy of the inverse of novelty. The two methods that achieved the best results using Tot Div were MF and UB. While MF was best for both MovieLens datasets, UB was superior for MovieTweetings. This result suggests that the density of the dataset affects the performance of the methods in terms of Tot Div, whereby MF can diversify results in dense datasets. However, in a sparse dataset, such as MovieTweetings, UB diversifies better than MF. When $\alpha\beta\gamma$ − Tot Div is used, both UB and MF are penalized, showing that robust performance in terms of Tot Div does not imply the same in terms of content novelty. Indeed, for this last measure, the difference between UB, MF, and IB diminishes to only one point in Ml100K and ML1M. The difference in favor of MF and UB increases with MovieTweetings. In general, these results reflect that $\alpha\beta\gamma - $ TOT DIV allows the evaluation with a single measure of both aspects of a recommender system method (i.e., shows the degree of novelty of the items in a given list and the extent to which the recommended lists are different).

When the dependency between the measures was tested, our hybrid measure $\alpha\beta\gamma$-TOT DIV shows that the results are consistently different from those obtained by the other measures. In addition, our results show that EFD, a measure provided in RankSys, is also independent from the other measures. Conclusions derived using EFD and $\alpha\beta\gamma$-TOT DIV are quite different. EFD indicates that the best method in terms of novelty is randomization, a conclusion that shows that randomization helps in novelty maximization, just as expected. When $\alpha\beta\gamma$-TOT DIV is used, the best methods are MF and UB, suggesting that these methods assist in genre diversification and in the reduction of the popularity-bias effect. Accordingly, insights derived from $\alpha\beta\gamma$-TOT DIV are not only different from EFD insights, but complementary.

In this article, we show that novelty in terms of item occurrence is only a partial view of the novelty concept. We believe that addressing a different feature of novelty, such as the proposed content novelty, will help to improve the way in which recommender systems are evaluated and, accordingly, reveal which methods are suitable for each purpose.

Acknowledgements Marcelo Mendoza was supported by Conicyt PIA/Basal FB0821 and the Millennium Institute for Foundational Research on Data.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

Abbassi, Z., Amer-Yahia, S., Lakshmanan, L., Vassilvitskii, S., Yu, C. (2009). Getting recommender systems to think outside the box. In *Proceedings of the 3rd ACM conference on recommender systems, RecSys* (pp. 285–288).

Alharthi, H., Inkpen, D., Szpakowicz, S. (2017). A survey of book recommender systems. Journal of Intelligent Information Systems.

Baeza-Yates, R., Hurtado, C., Mendoza, M., Dupret, G. (2005). Modeling user search behavior. In *Proceedings of the 3rd Latin American Web Congress, LA-WEB* (pp. 242–251).

Bellogin, A., Castells, P., Cantador, I. (2011). Precision-oriented evaluation of recommender systems: an algorithmic comparison. In *Proceedings of the 5th ACM conference on recommender systems, RecSys*.

Breese, J., Heckerman, D., Kadie, C. (1998). Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the 14th conference on uncertainty in artificial intelligence, UAI* (pp. 43–52).

Castells, P., Hurley, N., Vargas, S. (2015). Novelty and diversity in recommender systems. In Ricci, F., Rokach, L., Shapira, B. (Eds.) *Recommender systems handbook*. 2nd: Springer.

Channamsetty, S., & Ekstrand, M. (2017). Recommender response to diversity and popularity bias in user profiles. In *Procceedings of the 13th international florida artificial intelligence research society conference* (pp. 657–660).

Chen, H.-C., & Chen, A.L.P. (2005). A music recommendation system based on music and user grouping. *Journal of Intelligent Information Systems*, *24*(2), 113–132.

Clarke, C., Kolla, M., Cormack, G., Vechtomova, O., Ashkan, A., Büttcher, S., MacKinnon, I. (2008). Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st international ACM SIGIR conference on research and development in information retrieval* (pp. 659–666).

Dupret, G., Piwowarski, B., Hurtado, C., Mendoza, M. (2006). A statistical model of query log generation. In *Proceedings of the 13th international conference on string processing and information retrieval, SPIRE* (pp. 217–228).

Ekstrand, M., Tian, M., Madrazo, I., Ekstrand, J., Anuyah, O., McNeill, D., Pera, S. (2018). All the cool kids, how do they fit in? popularity and demographic biases in recommender evaluation and effectiveness. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 1–15).

Elberse, A. (2007). The power of stars: do star actors drive the success of movies? *Journal of Marketing*, *71*(4), 102–120.

Gomez, C., & Hunt, N. (2015). The netflix recommender system: algorithms, business value, and innovation. *ACM Transactions on Managements and Information Systems (TMIS)*, *6*(4), 1:19.

Hahsler, M. (2016). Recommenderlab: a framework for developing and testing recommendation algorithms, R package. https://CRAN.R-project.org/package=recommenderlab.

Hofmann, T. (2004). Latent semantic models for collaborative filtering. *ACM Transactions on Information Systems*, *22*(1), 89–115.

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of ir techniques. *ACM Transactions of Information Systems*, *20*(4), 422–446.

Järvelin, K., & Kekäläinen, J. (2000). Ir evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23rd international ACM SIGIR conference on research and development in information retrieval* (pp. 41–48).

Koren, Y., Bell, R., Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *IEEE Computer*, *42*(8), 30–37.

Kunaver, M., & Požrl, T. (2017). Diversity in recommender systems – a survey. *Knowledge-Based Systems*, *123*, 154–162.

Lops, P., de Gemmis, M., Semeraro, G. (2011). Content-based recommender systems: State of the art and trends. In Ricci, F., Rokach, L., Shapira, B. (Eds.) *Recommender systems handbook*. 1st: Springer.

Mendoza, M., & Baeza-Yates, R. (2008). A web search analysis considering the intention behind queries. In *Proceedings of the 6th Latin American Web Congress, LA-WEB* (pp. 66–74).

Pazzani, M. (1999). A framework for collaborative, content-based and demographic filtering. *Artificial Intelligence Review*, *13*(5-6), 393–408.

Pazzani, M., & Billsus, D. (2007). Content-based recommendation systems. In Brusilovsky, P., Kobsa, A., Nejdl, W. (Eds.) *The Adaptive Web*: Springer.

Popescul, A., Ungar, L., Pennock, D., Lawrence, S. (2001). Probabilistic models for unified collaborative and content-based recommendation in sparse-data environments. In *Proceedings of the 17th conference in uncertainty in artificial intelligence, UAI* (pp. 437–444).

Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., Riedl, J. (1994). Grouplens: An open architecture for collaborative filtering of netnews. In *Proceedings of the conference on computer supported cooperative work, CSCW* (pp. 175–186).

Saracevic, T. (1995). Evaluation of evaluation in information retrieval. In *Proceedings of the 18th international ACM SIGIR conference on research and development in information retrieval* (pp. 138–146).

Sarnè, G.M.L. (2015). A novel hybrid approach improving effectiveness of recommender systems. *Journal of Intelligent Information Systems*, *44*(3), 397–414.

Sarwar, B., Karypis, G., Konstan, J., Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international world wide web conference, WWW* (pp. 285–295).

Soboroff, I., & Nicholas, C. (2000). Collaborative filtering and the generalized vector space model. In *Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval* (pp. 351–353).

Takács, G., Pilászy, I., Németh, B., Tikk, D. (2007). Major components of the gravity recommendation system. *SIGKDD Explorations*, *9*(2), 80–83.

Trattner, C., Said, A., Boratto, L., Felfernig, A. (2018). Evaluating group recommender systems. In Felfernig, A., Boratto, L., Stettinger, M., Tkalcic, M. (Eds.) *Group recommender systems: an introduction* (p. 59): Springer.

Vargas, S. (2015). Novelty and diversity evaluation and enhancement in recommender systems. PhD thesis, Universidad Autónoma de Madrid, Spain.

Vig, J., Sen, S., Riedl, J. (2012). The tag genome: encoding community knowledge to support novel interaction. ACM Transactions on Intelligent Systems (TIST).

Wang, Z., Yub, X., Feng, N., Wang, Z. (2014). An improved collaborative movie recommendation system using computational intelligence. *Journal of Visual Languages & Computing*, *25*(6), 667–675.

Zhao, X., Niu, Z., Chen, W. (2013). Opinion-based collaborative filtering to solve popularity bias in recommender systems. In *Proceedings of the 24th international conference on database and expert systems applications, DEXA* (pp. 426–433).

Ziegler, C., McNee, S., Konstan, J., Lausen, G. (2005). Improving recommendation Ñists through topic diversification. In *Procceedings of the 14th international conference on the world wide web, WWW* (pp. 22–32).