How to compose product pages to enhance the new users' interest in the item catalog?

Nicollas Silva[1], Diego Carvalho[2], Adriano C. M. Pereira[1], Fernando Mourão[3], and Leonardo Rocha[2]

[1]Universidade Federal de Minas Gerais, Brazil
[2]Universidade Federal de São João del Rei, Brazil
[3]Seek AI Labs, Belo Horizonte, Brazil
{ncsilvaa,adrianoc}@dcc.ufmg.br
{dcarvalho,lcrocha}@ufsj.edu.br
fernando.mourao@catho.com

Keywords: Non-personalized RS, Pure Cold-Start problem, Users Coverage

Abstract. Converting first-time users into recurring ones is key to the success of Web-based applications. This problem is known as Pure Cold-Start and it refers to the capability of Recommender Systems (RSs) to provide useful recommendations to users without historical data. Traditionally, RSs assume that non-personalized recommendation can mitigate this problem. However, several users are not interested in consuming just biased-items, such as popular or best-rated items. Then, we introduce two new approaches inspired by user coverage maximization to deal with this problem. These coverage-based RSs reached a high number of distinct first-time users. Thus, we proposed to compose the product's page by mixing complementary non-personalized RSs. An online study, conducted with 204 real users confirmed that we should diversify the RSs used to conquer first-time users.

1 Introduction

Recommender Systems (RSs) have assumed a prominent role in Web-based applications, affecting decisively distinct business phases, such as the acquisition and retention of users. In the retention phase, the performance of current prediction models is extremely satisfactory [2]. A recent study highlighted RSs as the main responsible for 35% of sales on Amazon, 2/3 of the movies watched on Netflix and 38% more click-through on Google News [9]. However, the user acquisition phase has not received much attention in recent years. In this phase, RSs help to consolidate the users' first impression about the item catalog, which may influence the conversion rate of first-time users into clients [11].

In the literature, this problem is called Pure Cold-Start and it remains poorly exploited by researchers who just consider the Cold-Start problem [16]. Despite this, the Pure Cold-Start problem has grown in real domains since several users became to reach systems through incognito navigation or with social networks

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22734-0_24

disable due to privacy issues [20]. In this context, it is not easy to capture personal information from cookies, social networks or browsing history. For this reason, the users are always unknown, and the system always faces the challenge of recommending useful items for them who do not have any information [6].

In this work, we identify an opportunity for improvements on state-of-the-art non-personalized RSs that address the Pure Cold-Start. The literature assumes that items biased by popularity, recency or positive ratings are enough to attract first-time users. We show that a non-negligible portion of these users is not interested in consuming such items in some domains. Hence, exploiting biased-items RSs to compose product pages is not the best method to conquer distinct first-time users. This work aims to answer a promising research question: *How to compose product pages to attract the maximum number of first-time users?*

We hypothesize that to satisfy distinct first-time users, RSs should balance recommendations that suit distinct user profiles. Aiming to validate this hypothesis, we evaluated three state-of-the-art RSs and two novel strategies, proposed by this work. Traditional RSs are inspired by the utility of biased-items [14] - (1) *Most Popular*; (2) *Best-Rated*; and (3) *Recent Items*. We propose two novel non-personalized RSs inspired by user coverage maximization, already exploited to address other RSs related problems: (1) *Max-Coverage*: selects items that cover a large number of distinct users, such as addressed in [15]; and (2) *Niche-Coverage*: selects items that cover distinct user profiles [13]. Complementary of our last work [20], we propose an extension of the Niche-Coverage method and deeper analyzes than previous ones to consolidate their practical application.

Offline assessments on four popular datasets from e-commerce and entertainment domains evinced that the methods are complementary. While traditional RSs retrieved potentially relevant items, obtaining high utility, the new RSs enhanced diversity. Further, the new RSs reached a higher number of distinct first-time users. Therefore, mixing these complementary RSs to compose product pages is a promising answer for our research question in real scenarios. To confirm this assumption, we conducted an online study with 204 real users. We build an A/B test comparing traditional RSs (scenario A) against complementary RSs (scenario B). For each scenario, we asked the users to select movies of their interest and answer questions about the list of items. The results highlighted as main contribution a clear message: we should combine complementary non-personalized RSs in product pages.

2 Related Work

In the literature, the term Pure Cold-Start refers to a subtask of the Cold-Start problem [10]. Despite being closely related, both problems should be addressed differently. Whereas in the Cold-Start problem exists a lot of strategies to deal with small consumption history of users, in the Pure Cold-Start there are few strategies to handle first-time users [1]. We identified three main categories of RSs designed to deal with the Pure Cold-Start problem: (1) Knowledge RSs; (2) Social Filtering RSs; and (3) Non-Personalized RSs.

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22734-0_24

Knowledge RSs try to acquire user information using small questionnaires in user-web interaction. So, several studies have been proposed to improve the classical RSs with this information [21,5]. However, *He et al.* [5] argue that the quality of recommendations depends on information provided by users, who may not be able to define clearly their preferences. In turn, Social Filtering RSs exploit 'external' information about users, such as social or demographic data. In general, these RSs use hybrid methods to mitigate the Cold-Start problem [18,16]. Despite the advantages obtained, these approaches are not commonly used in e-commerce scenarios, because many users are not willing to provide demographic information before buying products.

Non-Personalized RSs are the predominant solution in real-world scenarios due to simplicity, domain independence, and efficiency. These RSs derive global information about items and users [2], exploiting key features related to consumption, such as popularity, ratings, and release/consumption recency [14]. However, these strategies are targeted to specific profiles, biasing users interested in items that satisfy a large portion of a population. To balance the recommendations for all users, the concept of result diversification has been introduced from the field of IR [23]. In general, the items recommended are re-ordered on the basis of a given diversification objective [22]. In this work, to attract more first-time users, we propose to diversify the items with user-coverage.

3 Handling First-Time Users

The Pure Cold-Start problem occurs when the system does not have any information about users. For this reason, first, we simulate these scenarios and, next, discuss the main approaches that address this problem.

3.1 First-Time Users Definition

First, we select the MovieLens 1M and 10M, and the CiaoDVD and Amazon datasets, described in Table 1, to simulate entertainment and e-commerce scenarios. Next, we simulate the first-time users in our datasets as follows. We sort the users considering the timestamp from the first item consumed in their historical data. Then, we selected the last 20% of users as the first-time ones, since they present the most recent actions in each collection. So, we used all historical data of the selected users to compose test sets and removed them from the training sets used as inputs by the evaluated RSs. The number of users selected from each dataset is available on the last column of Table 1.

Datasets	Users	Items	Sparsity	Genres	First-time
ML-1M	6,040	3,952	95.82%	18	1,277
ML-10M	69,878	10,283	98.60%	20	10,633
CiaoDVD	17,615	16,621	99.97%	17	3,523
Amazon	8,057	26,729	99.92%	471	1,612

Table 1. Datasets - general information.

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22734-0_24

3.2 Biased-Item Models

In Pure Cold-start problem, the state-of-the-art RSs are based on biased-items recommendations. These models assume that items biased by popularity, recency or positive ratings are useful to first-time users. For this reason, we implement and evaluate these non-personalized RSs, popularly used in real domains:

- **Popularity (Pop):** selects the k most popular items in the domain. The popularity is estimated by the number of distinct users who consumed an item i.
- **Best-Rated (BestR):** recommends the k best evaluated items in the domain. Basically, we sum the items' ratings and divide its by the number of users.
- **Recent Items (RecItems):** recommends the k last items consumed by users, calculated based on timestamp.

Generally, items recommended by these RSs are concentrated in the *head* of popularity distribution. However, several studies have discussed the *long tail* phenomenon in real scenarios such as Amazon and Netflix [7]. In these scenarios, tail products generate a significant fraction of the total revenue in aggregate and can boost head sales by offering consumers both their mainstream and specific tastes. For this reason, we suppose that there are many users interested in other items beyond the recommended by these state-of-the-art RSs. Hence, for every dataset, we select the top-100 items from Popularity, Best-Rated, and Recent Items, and count the number of biased-items in each user's consumption history. The values of each RS are normalized by the history size of each user and plotted in Figure 1. Values close to 100% indicate that user consumption is strongly biased by the items recommended and values close to 0% show that user consumption is formed by other items.

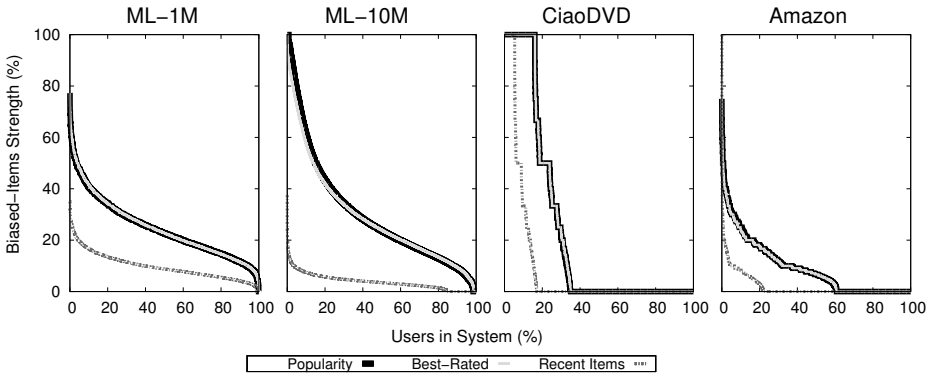


Fig. 1. Percentage of popular items consumed by all users.

In each ranking, we observe three user behaviors: (1) users who prefer biased-items (bias more than 70% - head of distribution); (2) users who prefer other items (bias less than 30% - tail of distribution); and (3) users who mix biased-items and others (bias around 30% and 70% - middle). These results show a

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22734-0_24

non-negligible portion of users with (2) and (3) behaviors, i.e., interested in other items beyond the selected by these RSs. Specifically, in the e-commerce domains, around 40% to 60% of the users do not have any biased-item in your consumption history. Therefore, these results point out an opportunity for improvements on state-of-the-art non-personalized RSs that address the Pure Cold-Start problem.

3.3 Coverage-based Models

Exploring the improvements opportunity, we propose two non-personalized RSs based on user coverage maximization. Max-Coverage is inspired by a NP-hard problem (Maximum k-Coverage), already exploited to address other RSs related problems [15]. In turn, we propose a new method, called Niche-Coverage, which aims to apply Max-Coverage in a distinct niche of users found by any clustering approach. Both methods consider that maximizing user coverage is a relevant approach to handle the Pure Cold-Start problem.

Max-Coverage: This strategy models the recommendation domains in sets of items and users, and applies the Maximum k-Coverage problem to find the items for first-time users. Formally, considering a universe of elements $U = \{u_1, ..., u_m\}$, a family of sets $F = \{S_1, ..., S_n\}$, where each set S_i is a subset of U and an integer k, the *Maximum k-Coverage* consists to find a subfamily $F^* \subseteq F$ such that $|F^*| \leq k$ and the number of covered elements $|\bigcup_{S \in F^*} S|$ is maximized, i.e. using up to k sets, cover as many elements as possible.

In a recommendations domain, we model the domain based on the users-items interaction, creating sets of users and items. Then, let $U = \{u_1, ..., u_m\}$ as the users that previously have consumed items, we create the set $S = \{S_1, .., S_n\}$, where each element S_i is a subset of users who consumed the item i. Therefore, the objective is to find the subset $S^* \subseteq S$, such that $|S^*| \leq k$ and the number of distinct covered users $|\bigcup S_i|$ is maximized. In another viewpoint, Max-Coverage is modeled as a bipartite user-item graph, where the nodes are the users and items, and the edges represent the interactions of a given user to an item. Then, MaxCov aims to select k items that reach the maximum number of distinct users, as proposed in other RSs related problems [15].

The Maximum k-Coverage is a NP-hard problem and there is no optimal solution in polynomial time. Our RS is a greedy algorithm to select the item that maximizes the number of users covered at each iteration. k iterations are executed to evaluate every set S_i was not selected (i.e., $S \in F \setminus F^*$). In each iteration, the algorithm looks for the item that maximizes the intersection of users not covered yet ($|S \cap R|$). A superficial analysis of this strategy can conclude that the selected items are the most popular ones, considering that the goal is to find items related to many users. However, at each iteration, the set R (resting users) is constantly updated to exclude users covered by the selected S-set ($R \leftarrow R \setminus S$). For this reason, this strategy recovers increasingly less popular items. The algorithm ends when k items are selected or when there are no more users to be covered. The complexity of this algorithm is $O(kmn)$, where k is the number of items to be recommended, m is the number of users and n is the number of items.

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22734-0_24

Niche-Coverage: This model is inspired by users' behavior studies [13]. Since the first surveys in RSs, the main approaches are often implemented using collaborative filtering (CF) algorithm [14,19]. CF algorithms produce recommendations based on the assumption that similar users have similar tastes. Then, people who share common ratings are a good source of recommendations. However, these algorithms are not able to Pure Cold-Start problem, because it is impossible to find similar users to first-time users. Nevertheless, the assumption used still true for our problem and it is the premise used by Niche-Coverage. In this case, our approach intends to divide users into niches of common interests and identify items that cover the most users for each niche. We suppose that recommending items from distinct niches of users, the system can reach all distinct preference of first-time users because we present the things that appealed to all types of users. First, we find the k items used to cover all users from a specific niche through the Max-Coverage algorithm. Next, we merge the items selected based on the size of each niche to maximize the number of users covered. In a recommendation list R of size k, the biggest niche compose the most of items in R.

The definition of users niches is based on clustering methods. In recommendations domains, the most famous clustering methods are the traditional *k-means* and *Bisecting k-means* [4]. These methods use the ratings assigned by users-items interactions to group users in sets with common interests (i.e., niche of users). In this work, we compare both clustering methods looking for the most suitable and the number of clusters to be used. Then, we should find the number c of clusters with Maximum Rate (CMR), oppositely to [13]. For this, we look for the number of clusters that maximizes the mean Hit Rate, a traditional metric of business performance often associated with sales [2]. Specifically, we are interested in the niche that maximizes the hit rate metric because is crucial that system shows at least one relevant item for users in this first interaction. This process is shown by the Equation $c = arg\ max \left[\sum_{n=1}^{N} \left(\frac{\sum_{u=1}^{U_{test}} |R_{list}(u) \cap I_{test}(u)|}{U_{test}}\right)\right]$, where N is the number of users niche, U_{test} the set of first-time users, $I_{test}(u)$ the items in test set consumed by u and $R_{list}(u)$ the recommendations generated by Niche-Coverage for the user u.

Therefore, the goal is to select the *representative items* from each niche of users, which are the items with the highest chances of matching the preference of any user from the niche. Initially, we classify the set of users U in c niches. Next, at each iteration, we analyze each niche of users. First, the set R is updated to contain only users from the niche evaluated. So, we select the subset S that maximizes the number of users covered. In this case, each element S_i in set $F = \{S_1, .., S_n\}$ is a subset of users from the cluster who consumed the item i. Next, we apply the Max-Coverage approach to find k items from each niche of users. Again, the set R (resting users) is constantly updated to exclude users covered by the selected S-set. A set of *Items* saves the k items selected for each niche. Then, finally, we execute a *merge* function to generate the final recommendation list. This function select items from each niche according to the Max-Coverage order. The complexity of this algorithm is divided into two steps, clusters computation, and Max-Coverage recommendation. The clustering complexity depends on the

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22734-0_24

implementation. In general, the complexity is $O(ncdi)$ where n is the number of d-dimensional vectors, c the number of clusters and i the number of iterations needed until convergence. However, in order to mitigate the Pure Cold-Start problem, we need to compute the clustering algorithm just one time, before the recommendation process. Hence, Niche-Coverage complexity is related to the recommendation step. Basically, it consists in to compute the Max-Coverage for c times (one for each niche of users). So, this complexity-time is $O(ckmn)$.

4 Empirical Assessments

This analysis aims to compare biased-items and coverage-based RSs for addressing the Pure Cold-Start problem. We used all historical data of the selected users in Section 3.1 to compose test sets and removed all data information about them. The other users compose the training sets and are used as inputs by the evaluated RSs. So, first, we analyze the best parameters to the Niche-Coverage algorithm, comparing k-means and Bisecting k-means. Next, we evaluated the recommendation lists issued by each RS, considering the most famous quality requirements. We also analyze the users reached by the items recommended, in order to consolidate the complementarity of our approaches. To attract first-time users with different preferences, it is not enough to assume that strategies focus only on the usefulness of items to users [7]. Aspects such as diversity, coverage, and surprise are important to compose an interface that presents the best of items catalog available to first-time users. The usefulness of each advisor is evaluated by *Hit Rate*, *Precision* and *Recall* [2]. The diversity of the recommended items is evaluated by the metrics of *ILD* and *Genre Coverage* [15].

4.1 Niche-Coverage Definitions

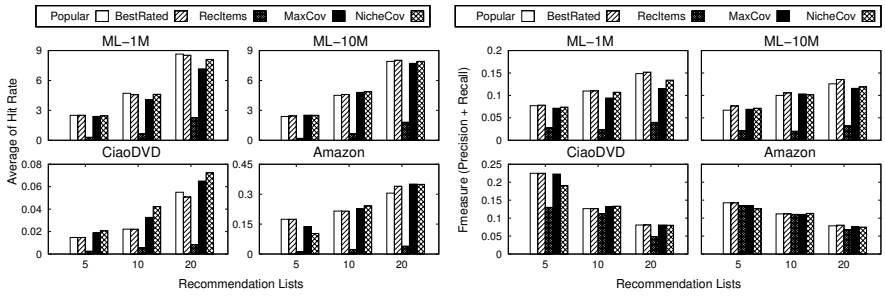
To define the best Niche-Coverage performance, we analyze two clustering methods and look for the number of clusters that maximizes the rate (CMR). Then, we compute 2^c clusters with the k-means and Bisecting k-means algorithms, where c range is $c = \{1, 2, 3, ..., 8\}$. Considering each number of clusters, we run the Niche-Coverage algorithm to recommend 10 items for first-time users and evaluate the Hit Rate metric. In the entertainment scenario, we find the best hit rate using k-means with 10 and 4 clusters, respectively in ML-1M and ML-10M. In the e-commerce domain, we find the best hit rate using k-means with 2 niches to CiaoDVD and Bisecting k-means with 93 niches in Amazon. Moreover, the results found are better than just using one cluster (i.e., the Max-Coverage approach). So, we confirm our premise that dividing users in niches and run the Max-Coverage locally is better than only run Max-Coverage with all users.

4.2 Quality of Recommendations

First, we simulate a real web-scenario, where users handle with 5, 10 and 20 items, and measure the RS's effectiveness. Then, we show the Hit Rate and F-measure metrics in Figure 2(a) and 2(b), respectively. In the entertainment scenario (ML-1M and ML-10M datasets), the users usually watch famous movies,

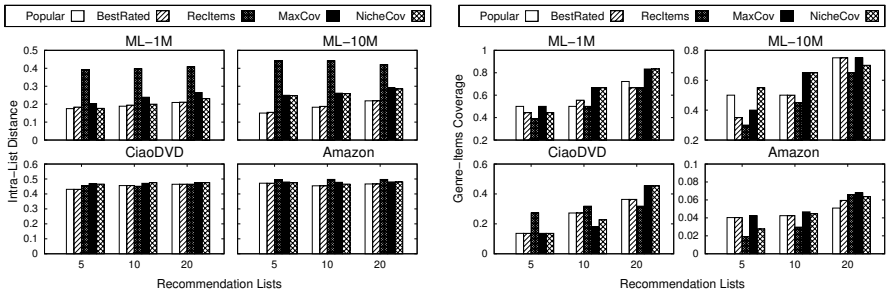
ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22734-0_24

that attracted the attention of many domain users. The Popularity and Best-Rated approaches have satisfactory performance in these scenarios. However, Max-Coverage and Niche-Coverage also have a high effectiveness rate. For example, in the ML-10M dataset, with 10 items recommended, the Niche-Coverage have the highest hit rate. In the e-commerce scenario, the users are interested in buy specific products, frequently related to their personal preference. For this reason, approaches based on biased-items are not the best option to satisfy first-time users. In this case, our Max-Coverage and Niche-Coverage approaches have the best performance, as shown in the second row of Figure 2. Specifically, in CiaoDVD dataset with just 10 items recommended, the Niche-Coverage has double the performance of state-of-the-art RSs. Statistically, we consolidated the results by Wilcoxon test for non-parametric distributions. In the entertainment scenario, the RSs' performance is not statistically different. In turn, in the e-commerce domains, the Niche-Coverage performance presents a statistical gain with 99% of confidence interval and $p\text{-}value = 0.01$.



(a) Hit-Rate (b) F-measure

Fig. 2. Results of utility metrics on all domains.



(a) ILD - Diversity (b) Genre Coverage

Fig. 3. Results of diversity metrics on all domains.

Furthermore, these gains obtained by our approaches are related to distinct items. Basically, due to the assumption of maximizing the coverage of users on the domain, Max-Coverage and Niche-Coverage recover items related to most of the users profiles. For this reason, these RSs are also high values of diversity and item-genre coverage, as shown in Figure 3(a) and 3(b). The Recent Items RS has the best value of diversity because it recommends just the last items consumed. However, these results are not efficient due to the low accuracy obtained

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22734-0_24

(Figure 2). On the other hand, the diversity presented by Max-Coverage and Niche-Coverage is achieved through potentially relevant distinct items. Specifically, in ML-10M dataset, our approaches have almost 50% of more diversity than traditional RSs with ILD metric. The same occurs in Genre Coverage metric, which Max-Coverage covers a greater number of distinct genres. We applied the Wilcoxon test, confirming the superiority of our approaches with a *p-value*= 0.001. These results show that Max-Coverage and Niche-Coverage are effective RSs, due to the high values of accuracy even gaining in terms of diversity.

4.3 Analysis of Complementarity

The last analysis point out to a complementary behavior between our approaches and the traditional ones. Our supposition is that Max-Coverage and Niche-Coverage do not recommend items biased by its rating or high influence. A straightforward analysis of the popularity of the first 10 items recommended by each strategy confirm this assumption. Our analysis demonstrates that all methods, except Best-Rated, recommend less popular items. The Max-Coverage recommends items that are less popular than the previous item. Niche-Coverage also diversifies the items from traditional RSs with less popular items. Hence, these analyses highlight that: (1) traditional approaches are much similar because they recommended biased-items; and (2) the new approaches are complementary to traditional methods because they recommended items based on the coverage.

Entertainment Scenario

	ML-1M			ML-10M		
RecLists	*top-5*	*top-10*	*top-20*	*top-5*	*top-10*	*top-20*
Popular	61.1% / **87.4%**	79.4% / **93.6%**	90.6% / **98.2%**	60.5% / **77.2%**	72.3% / **86.6%**	85.1% / **92.6%**
BestRated	64.1% / **88.1%**	78.7% / **93.8%**	90.4% / **98.1%**	62.0% / **82.1%**	73.1% / **87.0%**	84.7% / **92.7%**
RecItems	10.7% / **19.7%**	36.3% / **52.7%**	61.3% / **81.3%**	12.5% / **22.8%**	25.9% / **41.9%**	28.9% / **45.4%**
Max-Cov	69.6% / **89.3%** ▲	82.4% / **95.9%** ▲	91.9% / **99.3%** ▲	61.0% / **84.6%** ●	76.1% / **91.6%** ▲	86.6% / **96.3%** ▲
Niche-Cov	66.1% / **87.7%** ●	80.1% / **94.6%** ▲	91.3% / **98.9%** ●	62.9% / **84.6%** ●	76.6% / **91.5%** ▲	85.3% /**95.8%** ●

E-commerce Scenario

	CiaoDVD			Amazon		
RecLists	*top-5*	*top-10*	*top-20*	*top-5*	*top-10*	*top-20*
Popular	5.29% / **7.01%**	8.03% / **11.0%**	12.6% / **16.5%**	10.9% / **14.7%**	15.6% / **21.1%**	23.5% / **31.7%**
BestRated	5.29% / **7.01%**	8.03% / **11.0%**	12.6% / **16.5%**	10.9% / **14.7%**	15.6% / **21.1%**	24.5% / **31.9%**
RecItems	0.05% / **0.09%**	0.15% / **0.25%**	0.21% / **0.33%**	0.54% / **0.91%**	1.16% / **1.86%**	1.88% / **2.87%**
Max-Cov	5.56% / **7.43%** ●	8.92% / **11.7%** ●	13.1% / **17.3%** ●	10.9% / **14.7%** ●	16.3% / **23.0%** ●	25.6% / **34.4%** ▲
Niche-Cov	4.83% / **6.65%** ▼	8.50% / **11.2%** ●	12.9% / **17.1%** ●	8.12% / **12.0%** ▼	15.1% / **20.9%** ●	24.1% / **32.6%** ●

Table 2. This table shows the number of users conquered and covered by each RS, denoted by the set < *conquered/covered* >. The cells marked with a color ■ mean a higher number of users conquered by RS. The symbol ▲ denotes significant positive gains, ● non significant gains and ▼ significant negative losses. These gains are obtained concerning the best state-of-the-art RS, located in the first three rows, and applying a Chi-square test with 95% of a confidence interval.

However, in real web-scenarios, the system owners are interested in the user's satisfaction. If the users watch/buy their products, their profit will be higher. For this reason, we develop a metric to evaluate the number of users conquered by each RS, based on user satisfaction in real scenarios. We consider that a user is conquered by the system if s/he consumed and liked at least one item. In this work, we define that users like an item when they provide a rating greater than their personal average. We analyze each recommendation list, counting the number of users conquered by the items. Note that, this method is more complex than

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22734-0_24

simple coverage. We measure RS's coverage just considering if the user watched or bought the item recommended. Here, the ability to conquer is related to the rating assigned by users that watched or bought the item. In Table 2, we color the cases that the number of users conquered is higher than baselines and mark it with symbols of statistical significance. The Max-Coverage approach covers more users than other RSs due to its greedy algorithm. Moreover, we observe that in 9/12 cases, the Max-Coverage approach also conquers more users than baselines. In the other 3 cases, Niche-Coverage conquers more users.

We also count the number of users conquered exclusively by one RS. We observe in Table 3 that: (1) Popularity and Best-Rated do not aggregate users than those already conquered by other approaches; (2) Recent-Items conquers some different users, but it does not present items potentially relevant to first-time users; and (3) Max-Coverage and Niche-Coverage conquers more first-time users, which are distinct from others. These results point out a room for improvements, which are explored in the next section.

	Exclusive users of each RS			
RecList	**ML-1M**	**ML-10M**	**CiaoDVD**	**Amazon**
Popularity	1.19%	0.41%	0.00%	0.00%
BestRated	1.92%	0.84%	0.00%	0.00%
RecItems	2.20%	1.17%	0.14%	0.71%
Max-Cov	**3.84%**	**9.80%**	1.57%	2.91%
Niche-Cov	0.56%	8.74%	**1.72%**	**3.68%**

Table 3. Number of users conquered exclusively by one RS.

5 Construction of Product Pages

Mixing different recommendation lists on a product page is a common practice of real systems. However, we argue that these lists usually reach a similar subset of first-time users, since all of them are based on biased-items. To verify this behavior, we evaluate product pages composed by mixing three top-10 recommendation lists issued by distinct combinations of RSs. We restrict each page to have only three lists for working in smartphone scenarios, characterized by small screens. We evaluate five different combinations (Table 4). For each combination, we evaluate the number of users who rated positively at least one item from the three RSs, obtaining the percentage of *Users Conquered* for each combination. The results show that by mixing <BestR, MaxCov, NiCov> we can reach a high number of first-time users. In other words, in the evaluated scenarios, product pages should be composed by Best-Rated, Max-Coverage, and Niche-Coverage. This work suggests that systems incorporate our strategies to be used side by side, changing from the traditional approach for our suggestion.

	Users Conquered			
Approaches	**ML-1M**	**ML-10M**	**CiaoDVD**	**Amazon**
Pop, BestR, RecItems	87.35%	76.27%	8.10%	16.34%
Pop, BestR, MaxCov	88.99%	84.35%	9.60%	18.54%
Pop, BestR, NiCov	84.88%	83.75%	9.76%	19.33%
Pop, MaxCov, NiCov	87.25%	83.83%	9.76%	20.10%
BestR, MaxCov, NiCov	**89.15%**	**84.89%**	**9.76%**	**20.10%**

Table 4. Percentage of users conquered mixing three RS.

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22734-0_24

5.1 Online User-Centered Study

In order to evaluate our new approach for mixing RSs, we perform an experiment with volunteer users of different ages and preferences to evaluate the recommendations. Once the focus of this work is the first-time users, for who we do not have any information, a Web interface that presents the recommended items is able to simulate real scenarios. We follow the main guidelines of online evaluations presented in the literature [12]. We chose the movie scenario of *ML-Latest*, updated in August 2017. This dataset has 26M ratings assigned by 270K users to 45K movies on a scale from 1 to 5. The user-centered study was released during 8 days (from 09/07 to 09/14/2018), reaching 204 users that interacted with an online system. The users selected are 71% men and 29% women, from 11 to 63 years old. Moreover, 85% of users are frequent users of movies streaming systems. Initially, the participants are instructed to fill in a consent form. In the next three steps, users answer questions, selecting or ordering their favorite movies. In the end, the users answer questions about personal information. We are concerned in the three middle steps:

1. **A/B Test:** users have to choose one movie to watch or the option "None of the Movies". In this case, some users interact with a side A (traditional approach) and others with the side B (our approach).
2. **User Satisfaction:** users answer three questions about all movies presented in the first step. Basically, these questions are related to classical concepts, such as *unexpectedness*, *novelty* and *utility*.
3. **Ideal Ranking:** users have to build their ideal ranking between all movies presented in the first step. In this case, users can choose how many movies s/he wants. We suggest that users choose at least 5 movies.

The first step aims to compare side A (traditional approach) against the side B (our approach). Specifically, we present 10 movies of each RS, similarly to the current Web-scenarios. In this step, 102 participants interact only with the side A and the other half with the side B. Then, we ask for each user to select a movie to watch or the "None of the Movies" option. We are simulating real scenarios, where users have to make a decision: watch any movie or ignore the options. In this case, the labels have not a biasing effect because this step aims to highlight the most promising scenario instead of comparing the lists. Moreover, the users who interact with side A do not know about side B.

	Percentage of Users	
Recommendation list	**Side A**	**Side B**
Page top	39.21%	42.15%
Page middle	25.49%	17.64%
Page bottom	30.39%	33.33%
None	4.9%	6.8%

Table 5. User choices in the A/B test interface.

Table 5 shows the percentage of users that selected a movie from the list on top, middle or bottom of the page. In both sides, most users select movies from the list on the page top, related to Popularity (side A) and Best-Rated (side B). Despite the effectiveness of these RSs, this result may be related to the list

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22734-0_24

position on the page. However, there is a high percentage of users who roll down the page to select movies from the bottom lists, related to Recent Items and Niche-Coverage methods. Probably, this behavior is related to the complementarity of these RSs in the movies domain. This result reinforces the assumption that complementary RSs should be used to compose web pages.

The second step assesses the quality perceived by real users. In this step, we follow the questionnaire used in [17]. We present each movie with a short synopsis and ask users: (1) *Did you already know about this movie before this recommendation? (Yes, No, Don't know)*; (2) *Have you already watched this movie? (Yes, No)*; and *Would you like to watch this movie (for the first time or once again)? (Yes, No, Don't know)*. We create a ranking with users answers. In the first question, we count the number of users who said "no" to simulate a feeling of surprise by something *unexpected*. In the second, we count users who said "no" to measure the RS *novelty*. For the third question, we count the users who said "yes" to discover the RS *utility*. Table 6 summarize these rankings with the area under the curve (AUC) normalized by its highest value.

	Real User Satisfaction		
Recommender	**Unexpected**	**Novelty**	**Utility**
Popular	0.1929	0.4162	0.7196
BestRated	0.2570	0.5043	0.7688
RecItems	0.4798	0.6269	0.7177
Max-Cov	0.3394	0.4776	0.6647
Niche-Cov	0.3394	0.4912	0.7049

Table 6. Quality perception of real users.

Indeed, all five RS are useful for real users in web-scenarios. Conversely the offline results, Recent Items is also useful for real users because it recommends distinct items from other RSs. Moreover, Max-Coverage and Niche-Coverage also recommended unexpected items, increasing novelty for users. These results are reinforced in the next step. Specifically, the third step aims to compare each recommendation list with the ideal ranking built by users. We use a traditional pooling strategy from the IR field to create a ground-truth about users. Basically, we select top-10 results from each RS, removing the items duplicates and present these movies to participants in a random order. The users have to order the movies according to their preference. Then, we can measure three ranking metrics to compare the recommendations and the feedback provided by the user: *Jaccard Similarity*; *Mean Reciprocal Rank (MRR)* [3]; and *Normalized Discounted Cumulative Gain (nDCG)* [8]. These metrics measure, respectively: (1) the similarity of each recommendation list with the ideal ranking; (2) the position of the first relevant item recommended; and (3) the effectiveness of each RS.

	Ranking Metrics		
Recommender	**Jaccard**	**MRR**	**nDCG**
Popular	**0.35**	0.32	**0.49**
BestRated	**0.35**	0.27	0.46
RecItems	0.02	0.04	0.03
Max-Cov	0.33	0.32	0.47
Niche-Cov	**0.35**	**0.33**	**0.49**

Table 7. Quality of RSs based on users' ranking.

Table 7 shows the metric's average for 118 participants that built their ground-truth. We consider only the five-first movies to create a fair analysis.

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22734-0_24

The first column shows the Jaccard's similarity and does not highlight any difference between the rankings. The reason for this result may be there are a few distinct items to be selected in this step. In turn, the MRR shows a higher value to Max-Coverage and Niche-Coverage than baselines approaches. In other words, these RSs are more likely to present relevant items in the first positions than the other RSs. In addition, *nDCG* metric confirms the effectiveness of our RSs, comparing with the baselines approaches. Thus, this user-centered study points to improvement possibilities for owners of web-applications.

6 Conclusions

Web applications assume that items biased by popularity, recency or positive ratings are enough to suit most of the first-time user's profiles. However, this work shows that it is not always true because there are many users not interested only in biased-items. Conversely, we introduce two new methods inspired in user coverage maximization: Max-Coverage and Niche-Coverage. While traditional RSs retrieved potentially relevant items, obtaining just high accuracy, the new RSs keep accuracy and enhance diversity. Our experiments show a statistical gain in both concepts. Further, the new RSs match the interest of a higher number of distinct first-time users. Thus, these results highlight complementary behaviors between our RSs and traditional approaches and show an opportunity for improvements to compose product pages. We assume that to enhance the interest of first-time users on the item catalog, the web-applications should mix these complementary RSs. An online user-centered study with 204 participants reinforces this assumption with metrics related to user satisfaction.

Acknowledgments

This work was partially funded by the INWeb (no. 573871/2008-6), MASWeb (FAPEMIG/PRONEX APQ-01400-14), CAPES, CNPq, Finep, and Fapemig.

References

1. I. Barjasteh, R. Forsati, F. Masrour, A.-H. Esfahanian, and H. Radha. Cold-start item and user recommendation with decoupled completion and transduction. In *Proceedings of the 9th ACM RecSys*, pages 91–98, 2015.
2. J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez. Recommender systems survey. *Knowledge-Based Systems*, 46:109–132, 2013.
3. O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM CIKM*, pages 621–630, 2009.
4. M. A. Ghazanfar and A. Prügel-Bennett. Leveraging clustering approaches to solve the gray-sheep users problem in recommender systems. *Expert Systems with Applications*, 41(7):3261–3275, 2014.
5. C. He, D. Parra, and K. Verbert. Interactive recommender systems: A survey of the state of the art and future research challenges and opportunities. *Expert Systems with Applications*, 56:9–27, 2016.

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22734-0_24

6. A. Hernando, J. Bobadilla, F. Ortega, and A. Gutiérrez. A probabilistic model for recommending to new cold-start non-registered users. *Information Sciences*, 376:216–232, 2017.
7. Y.-C. Ho, Y.-T. Chiang, and J. Y.-J. Hsu. Who likes it more?: mining worth-recommending items from long tails by modeling relative preference. In *Proceedings of the 7th ACM WSDM*, pages 253–262, 2014.
8. K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.
9. D. Lee and K. Hosanagar. Impact of recommender systems on sales volume and diversity. 2014.
10. B. Lika, K. Kolomvatsos, and S. Hadjiefthymiades. Facing the cold start problem in recommender systems. *Expert Systems with Applications*, 41(4):2065–2073, 2014.
11. A. Majumdar and A. Jain. Cold-start, warm-start and everything in between: An autoencoder based approach to recommendation. In *IEEE IJCNN*, pages 3656–3663, 2017.
12. J. O'Donovan, N. Tintarev, A. Felfernig, P. Brusilovsky, G. Semeraro, and P. Lops. Joint workshop on interfaces and human decision making for recommender systems. In *Proc. 9th ACM RecSys*, pages 347–348, 2015.
13. A. L. V. Pereira and E. R. Hruschka. Simultaneous co-clustering and learning to address the cold start problem in recommender systems. *Knowledge-Based Systems*, 82:11–19, 2015.
14. A. Poriya, T. Bhagat, N. Patel, and R. Sharma. Non-personalized recommender systems and user-based collaborative recommender systems. *Int. J. Appl. Inf. Syst*, 6(9):22–27, 2014.
15. S. A. Puthiya Parambath, N. Usunier, and Y. Grandvalet. A coverage-based approach to recommendation diversity on similarity graph. In *Proceedings of the 10th ACM Conference on Recommender Systems*, pages 15–22. ACM, 2016.
16. A. N. Rosli, T. You, I. Ha, K.-Y. Chung, and G.-S. Jo. Alleviating the cold-start problem by incorporating movies facebook pages. *Cluster Computing*, 18(1):187–197, 2015.
17. M. Rossetti, F. Stella, and M. Zanker. Contrasting offline and online results when evaluating recommendation algorithms. In *Proceedings of the 10th ACM Conf. on Recommender Systems*, pages 31–34. ACM, 2016.
18. L. Safoury and A. Salah. Exploiting user demographic attributes for solving cold-start problem in recommender system. *Lecture Notes on Software Engineering*, 1(3):303, 2013.
19. S. Sedhain, S. Sanner, D. Braziunas, L. Xie, and J. Christensen. Social collaborative filtering for cold-start recommendations. In *Proceedings of the 8th ACM Conference on Recommender systems*, pages 345–348. ACM, 2014.
20. N. Silva, D. Carvalho, A. C. Pereira, F. Mourão, and L. Rocha. The pure cold-start problem: A deep study about how to conquer first-time users in recommendations domains. *Information Systems*, 80:1–12, 2019.
21. H. Steck, R. van Zwol, and C. Johnson. Interactive recommender systems: Tutorial. In *Proceedings of the 9th ACM RecSys*, pages 359–360, 2015.
22. 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*, pages 209–216. ACM, 2014.
23. S. Vargas, P. Castells, and D. Vallet. Intent-oriented diversity in recommender systems. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 1211–1212. ACM, 2011.

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22734-0_24