# When Diversity Is Needed... But Not Expected!

Sylvain Castagnos, Armelle Brun, Anne Boyer

▶ **To cite this version:**

Sylvain Castagnos, Armelle Brun, Anne Boyer. When Diversity Is Needed... But Not Expected!. International Conference on Advances in Information Mining and Management, Nov 2013, Lisbon, Portugal. pp.44-50. hal-00931805

**HAL Id: hal-00931805**
**https://inria.hal.science/hal-00931805**

Submitted on 15 Jan 2014

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# When Diversity Is Needed... But Not Expected!

Sylvain Castagnos*, Armelle Brun* and Anne Boyer*
*KIWI team - LORIA
Campus Scientifique, B.P. 239
54506 Vandœuvre - France
Email: {sylvain.castagnos, armelle.brun, anne.boyer}@loria.fr

***Abstract*—Recent studies have highlighted the correlation between users' satisfaction and diversity within recommenders, especially the fact that diversity increases users' confidence when choosing an item. Understanding the reasons of this positive impact on recommenders is now becoming crucial. Based on this assumption, we designed a user study that focuses on the utility of this new dimension, as well as its perceived qualities. This study has been conducted on 250 users and it compared 5 recommendation approaches, based on collaborative filtering, content-based filtering and popularity, along with various degrees of diversity. Results show that, when recommendations are made explicit, diversity may reduce users' acceptance rate. However, it helps increasing users' satisfaction. Moreover, this study highlights the need to build users' preference models that are diverse enough, so as to generate good recommendations.**

***Keywords*—Recommender systems, diversity, user modeling, user study.***

## I. Introduction

Recommender systems aim at helping users during their information search, by suggesting items that fit their needs and preferences. Recommender systems, that have emerged two decades ago, have been much studied by academic researchers, and are now an indispensable part of most of web services. However, a paradox is remaining in recommender systems: most of recommender systems aim at maximizing the precision of the recommendations, but do not consider human factors, which have an important role in decision processes. For example, in 2009 the Netflix Prize [1] has been won by BellKor's Pragmatic Chaos team, after a three-year long competition. The mean quadratic error (RMSE) has been improved by two hundredth [2], [3]. However, the corresponding algorithm has never been used, as it has become obsolete due to the emergence of new interaction modes and new user behaviors [4].

During the same period, works focusing on users' acceptance and adoption of recommender systems have shown that a difference of 10% of the RMSE cannot be perceived by users [5], [6]. However, these studies highlight the influence of human factors on users' satisfaction. These factors can be users' confidence in the recommender, the explanations provided by the recommender or the need in diversity of recommendations. The work conducted in this paper focuses on this last factor: the diversity of the recommendations. The experiments conducted in 2010 by Castagnos *et al.* [5] aimed at identifying the steps of a user's decision process, when facing a recommender system. In that study, diversity, which has not been anticipated, has appeared as an important factor in decision processes, but the experiments conducted did not allow to quantify the importance of this factor.

The contributions of our work are the following: we conduct a new user study, that aims at better understanding the role and the impact of diversity in recommender systems; several recommender systems algorithms (Collaborative Filtering (CF), Content-Based Filtering (CBF) and Popularity-based filtering (POP)) are implemented, as well as two new hybrid algorithms (combining CF and CBF), that allow to tune the degree of diversity of the recommendations. Analyzing these last two algorithms relies on a five-level inter-group model (one group for each algorithm) and a two-level intra-user model (implicitly and explicitly provided recommendations) in the movie domain. The training dataset is made up of a complete description of more than 500 movies and more than 3000 users. The experiment has been conducted during a one-week period, where data about 250 users has been collected. These users have been randomly split in groups of 50 users.

This study confirms the positive influence of the diversity on users' satisfaction. It also surprisingly highlights the importance of building preference models with enough diversity between items, especially for the cold-start phase, by encouraging users to rate different items. In addition, this study shows that despite having a positive impact on users' satisfaction, diversity has to be used carefully, as too much diversity may result in users who do not understand the coherence of the recommendations provided.

This paper is organized as follows: Section II is an overview of the state of the art of diversity in recommender systems, from conception and evaluation points of view. Section III is dedicated to the presentation of our study and Section IV presents and discusses the results. The last section concludes this paper.

## II. Related work

Diversity is an emerging research topic in recommender systems. Diversity is well known for playing an important role in the improving the interaction between users and information retrieval systems [7], [8]. However, the question about knowing why and how to improve the diversity remains open. Two main approaches are adopted in the literature. The first one analyzes the impact of diversity on users' behavior. The second one integrates diversity in machine learning algorithms from recommender systems. Both following subsections present these approaches.

### A. Role and impact of diversity

Diversity in recommender systems has been defined by Smyth and McClave [9] as the opposite dimension to similarity. We choose to refine this definition by defining diversity

as the measure that quantifies the dissimilarity within a set of items. Thus, the task of introducing diversity in a recommender system consists in finding the best set of items that are highly similar to users' known preferences by taking care to not freeze recommendations (if novelty is never introduced in recommendations) as well as taking care to recommend sets of items not too similar. The first case is referred to as intrinsic diversity, which avoids redundancy between the items to be recommended [10]. The second case is referred to as extrinsic diversity, which aims at alleviating the uncertainty due to data ambiguity or sparsity in user preference models, by recommending a large set of items [11]. In both cases, mechanisms used to introduce diversity rest on the same metrics (see section II-B). Note that a new classification of diversity has been recently proposed by Adomavicius and Kwon [12]. It distinguishes individual diversity and aggregated diversity, depending on if we are interested in generating recommendations to individuals, or to groups of users. Here, we focus on individual diversity.

The seminal work focusing on the role of diversity in recommender systems has been conducted on conversational recommender systems [7]. It has been the first to show that diversity improves the efficiency of recommendations. Works presented by Zhang and Hurley [13], or Lathia [14], even talk of user frustration when no diversity is provided. McGinty and Smyth [7] have also put forward the issues related to this dimension. For example, diversity does not have to be integrated in each recommendation step.

To thoroughly understand this last point, we focus on the two steps of an item selection process in information access systems [15]. In the first step, the user uses the system's interface to identify the pertinent criteria for his/her current search. The identification of these criteria is made, most of the time, by trial and errors, the recommendation cycles. The second phase aims at comparing all the possible solutions related to these criteria. Users unconsciously use a trial and error approach, by choosing an item, having a look to the related recommendations, then making a backward step. Once a first starting point is found, he/she continues his/her exploration of the set of items by using the recommendations (as well as the recommendations related to the recommendations chosen, etc.).

Given this behavioral model, we can admit that, as presented by McGinty and Smyth [7], diversity does not positively impact each recommendation cycle, especially in those where the user aims at increasing his confidence in the system. This conclusion is confirmed by Castagnos [5], which measured the evolution of diversity need through time.

Many discussions about the role of diversity have emerged these last years. McNee *et al.* [16] studied the limitations of precision measures used in recommender systems: less accurate recommendations may be more pertinent from the users point of view. They also focused on the difference between the diversity proposed to regular users and to new users. Several works also highlight that diversity is intrinsically present in collaborative filtering-based recommendations, through serendipity [17], [18]. In parallel, some works focused on the best way to present the recommendation, so as diversity is perceived by users [19], [20].

### B. Integrating diversity in recommender systems

The design of a recommender systems can divided into thee parts: (1) implicit or explicit collection of user traces, left by users when interacting with the system (preferences, tastes, usage, context); (2) building user models, based on these traces; (3) exploiting these models and machine learning algorithms to determine the adequate set of recommendations to be proposed to the active user.

Recommender systems are generally split into two families [21]: collaborative and content filtering. Until recently, the approaches dedicated to the improvement of the diversity were developed in the frame of content filtering [22], [23], [24]. These mechanisms can be used directly on the metric, or/and on the clustering/ranking algorithm used to generate recommendations. These approaches aim at increasing the diversity at the level of the attributes of the items.

In [9], diversity is represented by several metrics, that rely on the similarity between items: the more the items are similar, the lower is the diversity between them. Similarity between two items is defined as the weighted sum of the similarities on the attributes (see (1)).

$$Similarity(i_1,i_2) = \frac{\sum_{j=1..n} w_j * sim_{attribute=j}(i_1,i_2)}{\sum_{j=1..n} w_j} \quad (1)$$

Starting from this similarity metric, Smyth and McClave [9] has introduced two new diversity measures. The first one, called $Diversity$, computes the average dissimilarity within a class $C$, made up of $m$ items. The second one is a relative diversity ($RelDiversity$), that computes the added value in terms of diversity of an item on a class of items $C$ (see (2)).

$$RelDiversity(i,C) = \begin{cases} 0 \text{ if C = \{\},} \\ \frac{\sum_{j=1..m}(1-Similarity(i,c_j))}{m} \text{ otherwise.} \end{cases} \quad (2)$$

These metrics have then been used in content-based filtering to reorder the recommendation list, according to a diversity criterion. Two main approaches have been proposed: clustering-based [25] and selection-based approaches [22]. In clustering-based approaches, the aim is to build an optimal class of items, compared to a diversity criterion (that corresponds to the maximal diversity). The selection-based methods integrate the diversity in the recommender systems, without decreasing precision. Bradley and Smyth have been the first to propose a greedy-based selection algorithm, to find the most similar items to a user query, which are also diverse by pairs [22]. This algorithm selects the $K$ most similar items to a target item $t$ (see (1)). The recommendation list is filled iteratively, by choosing at each step the best quality item (see (3)), until getting the $top - N$ recommendations ($K < N$).

$$Quality(i,t,C) = Similarity(i,t) * RelDiversity(i,C) \quad (3)$$

The $top-N$ reordering algorithms, as in [22], are known for their tradeoff between speed and accuracy (including precision and diversity). Radlinski *et al.* propose 3 methods that rely on query reformulation, in order to increase diversity in the $top-N$ list [11]. Zhang and Hurley suggest to maximize the diversity while not decreasing the similarity; they view this task as a binary optimization problem [13].

In addition to these content-based algorithms, works have focused on a way to integrate diversity in collaborative filtering. Ziegler and McNee have proposed a generic formalism based on an intra-list similarity (ILS) and a $top-N$ selection, which can be used in several algorithms, such as collaborative filtering [26]. Said *et al.* [27] have studied a new way to integrate diversity in collaborative filtering, by adapting the clustering algorithms. These 3 last works are purely based on collaborative filtering, thus use similarity measures based on votes, not on attributes.

### C. Discussion

The main goal of this paper is to understand thoroughly the impact and the utility of diversity during the interaction between users and recommender systems. The state of the art presented in section II-A has illustrated how complex this dimension is, without having a complete view of all its dimensions. Indeed, the studies conducted to this date have measured the impact of diversity on satisfaction *a posteriori*, by using questionnaires [6]. Even if many studies have measured the impact of diversity on users' satisfaction with content-based filtering [7], [13], [14] on one side or collaborative filtering [28], [29] on the other side, no user study has been conducted to understand the role of diversity by comparing these two families of algorithms. [17] has addressed the diversity dimension thanks to the serendipity of such algorithms, but the degree of diversity is not controlled, and not always guaranteed. In this paper, we thus propose to conduct a user study, that focuses on diversity an that allows to: compare different families of algorithms (collaborative filtering, content-based filtering, popularity-based filtering); study users' behavior: from the collection of traces to the choice of items to be recommended, and check if diversity only plays a role during the recommendation phase as suggested by the literature, or if it also impacts the process of user modeling; study users' perception of diversity and differences in users' behavior, when recommendations are implicitly or explicitly presented.

Section II-B has shown that no hybrid algorithms (collaborative and content-based) that allow an equilibrium between accuracy and diversity of recommendations exist. To cope with this lack, we took inspiration from [22] and [26] to design two new algorithms, that combine collaborative and content-based filtering, to get the desired level of diversity. We propose to compare five algorithms.

## III. Experiment Setup

### A. Support

We conducted our experiment in the domain of cinema for several reasons. First, it is quite easy to collect a large dataset related to movies, so as to observe users' behaviors in a realistic context. Second, movies have a great number of attributes and are rated very often by users, unlike some other types of items. At last, cinema is a popular domain users are familiar with. This maximizes the chances that users know enough items in the proposed lists.

For the needs of our experiment, we built a website [30] and paid attention to users' cognitive load by spreading the experiment on several pages.

We started by collecting as much data as we could about the content of more than 500 movies, which includes titles, summaries, pictures, trailers, average ratings of press and spectators (and the corresponding number of ratings), movie genres, actors and actresses, directors, writers, release dates, languages, runtimes, and the fact that they belong to a saga or not. This information allow users to recognize movies, and is used by our recommender system as a training set to implement content-based filtering algorithms.

Collaborative filtering requires individual ratings from a large number of users. For this reason, we collected ratings from 3,158 users for the training data. In order to do so, we first gathered all the ratings of Allociné's real users [31] for the 509 selected movies. Then, we cleaned the database so as each user provides at least 20 movie ratings, and each movies is rated by at least 20 users. These thresholds represent the minimum number of ratings estimated by [32] to reach a good recommendation precision and quality level with collaborative filtering algorithms. We had to build this training set by our own, instead of relying on MovieLens or NetFlix corpuses, so as to guarantee that it is always possible to compute similarities between movies whatever are the attributes used. The size of our corpus is quite comparable to MovieLens. Moreover, and contrary to MovieLens, we have provided a good distribution of movies in term of popularity. We selected movies by paying attention to the fact that they must have more than 200 ratings on IMDb [33], and we manually checked with a sample group of 20 users that all the selected movies are known by most of them. Likewise, we randomly selected movies among those matching our criteria, while ensuring a good distribution on the rating scale from 1 to 5 (and in particular a good representativeness among the top-250 and the bottom-100 of IMDb movies). The average rating from the 3,158 users of the training set is 3.66, with a standard deviation of 1.37. To summarize, the whole set of information on movies and ratings represents our training dataset. The latter has been created thank to APIs of IMDb and Allociné. Characteristics of this dataset are made explicit in Tab I.

TABLE I: FEATURES FROM THE TRAINING DATA

| Type | Movies | Actors | Directors | Writers | Genres | Countries | Sagas | Ratings |
|---|---|---|---|---|---|---|---|---|
| Number | 509 | 903 | 310 | 351 | 23 | 17 | 98 | 173,120 |

### B. Algorithms

We used 3 algorithms from the state-of-the-art, called **POP**, **CBF** and **CF**. We also propose 2 new hybrid algorithms called **CFRD** and **CFFD**. The choice and the implementation of these algorithms have been motivated by a need of personalization in real time. We consequently had to choose algorithms that are known to be fast and precise, and adapted them to our architecture.

All the pieces of information provided by the volunteers of our user study constitute the test set. Recommendations are computed from the active user's profile and the training set. None of the data from the test set is included in the training set (i.e., our 250 users are different from those in the training set). In this way, recommendations are computed in the same conditions for all the 250 volunteers of this study.

**POP** is our baseline and recommends items randomly chosen among most popular items.

**CBF.** This algorithm recommends items in function of their similarities with items liked by the active user. In this case, we voluntarily focus on preference similarity, rather than recommendation diversity, to verify if users perceive a difference in comparison with other algorithms.

So as to compute the similarity between two movies (see (1)), we optimize weighting coefficients and similarity measures per attribute on our training set. The weighting coefficients on the different attributes are: $w_{date} = 0.5$ ; $w_{director} = 1$ ; $w_{actor} = 1$ ; $w_{genre} = 1.5$ ; $w_{language} = 0.25$ ; $w_{popularity} = 0.5$ ; $w_{saga} = 1$ ; $w_{scenarist} = 0.25$. Thus, as an example, the fact that two movies have the same director has two times more impact in the similarity computation than the fact that the released dates are closed from each other.

Similarity measures per attribute are defined as: $sim_{actor}(i_1, i_2) = \frac{\cap_{actors}}{\cup_{actors}}$ ; $sim_{genre}(i_1, i_2) = \frac{\cap_{genres}}{\cup_{genres}}$. The similarity for the release date is equal to 1 if the gap between the release dates is less than 5 years, 0 otherwise. The similarity for the popularity is equal to 1 if the two movies belong to the same popularity class (when the difference between the average ratings of these two movies is below a fixed threshold, and when the numbers of ratings for each of these movies are quite comparable). At last, similarities for the director, language, saga and writer are equal to 1 if the two movies have the same value for this attribute, 0 otherwise.

**CF.** We used an item-based collaborative filtering algorithm, as proposed in [34]. This algorithm transforms the user-item rating matrix in an item-item similarity matrix. Then, it applies a formula to predict the rating of an item $i$ that has not been rated by the active user yet. This rating is the mean of the ratings already provided by the active user, weighted by the similarities between the item $i$ and each of the items contained in the preference model of the active user. Our implementation relies on the Pearson correlation coefficient. At each iteration, we select the 10 items that got the highest predicted notes.

At last, we conceived two new algorithms called CFRD and CFFD, variants of the CF algorithm with a content-based hybridation. Our objective was to make the diversity level vary within the recommendation set. These two alternatives allow us to study the possible differences of users' perception when confronted with different diversity levels.

**CFRD.** (Collaborative Filtering with Relative Diversity). This algorithm first applies CF algorithm to compute the top-50. The first element of top-50 is included in the recommendation set. Then, items are added one by one, by selecting at each iteration the item from the top-50 that maximizes the relative diversity, in comparison with items already in the recommendation set (eq. (2)). We continue until we reach the expected number of recommendations.

Let us notice that this algorithm is quite similar to the algorithm proposed by [26], which is re-used in several user studies [28], [29]. In these papers, they build the top-10 recommendations by re-ranking the top-50 items of a CF algorithm according to a diversity metric. However, in our case, we used a different CF algorithm and a more complete diversity metric. In [29], authors explain that their diversification algorithm only reduces the similarity between movies in terms of genre and recognize that this may not fit the definition of similarity as the users of the system judge it.

But, more importantly, [26] use a diversification factor to find a compromise between the ranking of the CF algorithm and the diversity-based ranking. In other words, the higher movies are in the CF ranking, the more they have chances to appear in the top-10 list of the diversification algorithm. In our case, we consider that all of the movies in the top-50 of the CF algorithm are relevant, and have the same level of importance (except for the first one). Thus, we only re-rank the top-10 according to our diversity metric. In this way, we can more easily measure the impact of our diversity-driven approach, since diversity has more weight in the re-ranking phase.

**CFFD.** (Collaborative Filtering with Fixed Diversity). It is quite similar to CFRD, except that only a fixed percentage (x%) of recommendations has to come from the CF algorithm. In other words, instead of initializing the recommendation set with the first element of the top-50 (CFRD), we select the $n$ first items of the top-50 (with n = the expected number of recommendations * x%). In our implementation, this threshold has been fixed to 60%.

### C. Procedure

Our experiment is expected to last from 15 to 20 minutes per participant. After a short homepage that introduces the context of our study, each volunteer is invited to complete the 4-step procedure described below.

**Step 1.** A first questionnaire allows us to collect demographic data (first name, last name, email, gender, age, nationality, profession), and users' habits related to cinema (frequencies of visits in theaters, movie genres that they like, with whom they go to theaters and how they choose a movie, and if they read websites, magazines or books related to movies). Those habits reveal the users' expertise level in the domain of cinema. These questions are only uses for statistics on participants, and eventually to discard users whose answers might be irrelevant. At the end of step 1, each user is registered and the system assigns him/her one of the 5 recommendation algorithms. As a consequence, the participants are automatically spread into 5 groups. Each group has the same number of users.

**Step 2.** The system asks each user to rate a set of 100 films from 1 (I hate) to 5 (I like), under the pretext of filling his/her preference model. These 100 movies are displayed 10 by 10 (on 10 different pages), to avoid fatigue and cognitive overload. By default, movies are displayed in a synthetic way with minimal information such as the title, movie cover, director, genres, main actors and release date. However, participants

can get more details from the interface. What users do not know is that only the 3 first pages of ratings (30 movies) are common to every participant to initialize profiles. In pages 4 to 10, the list of movies to be rated is provided by the recommendation algorithm that has been assigned to the active user. In this way, movies are likely to interest users, but they are not aware of the fact that these lists are built in accordance with their preferences. To avoid any bias, movies are displayed in a random order on each page. Of course, given the size of our training set (509 movies) there is a risk that the quality of recommendations decreases due to the lack of interesting but not rated items. However, this risk is low since they only rate 20% of the database. Moreover, this phenomenon impacts all the algorithms in the same way.

**Step 3.** The system proposes a one-week TV program (one movie for each day of the week). The TV program is made of five TV channels, one for each recommendation algorithms (POP, CBF, CF, CFRD, CFFD). In this phase, we explicitly told users that these channels were made of recommendations from different algorithms. The goal was to measure if there are differences of confidence level toward these algorithms, if users can distinguish the recommended lists, and if the lack of diversity can pose a problem over a week. Users have to order channels from 1 to 5 according to their preferences.

**Step 4.** A post-questionnaire allows users to make explicit and quantify the performance of algorithms during step 3. In particular, we asked users to evaluate the recommendation relevance, the diversity levels during steps 2 and 3, and their confidence level within their ranking of step 3. We used a Likert scale with 7 modalities.

### D. Hypotheses

Before conducting this experiment, we enumerated the following hypotheses:
**H1.** Users perceive diversity. Results from the post-questionnaire should reflect this tendancy, in particular for groups assigned to CFRD and CFFD.
**H2.** Diversity improves users' satisfaction. Ratings collected in Step 2 should be higher for groups assigned to CFRD and CFFD, than for the other groups.
**H3.** Content-based algorithms increase the level of confidence of users, on the contrary of those based on diversity. Recommendations from the CBF algorithm should meet with more success in Step 3.

### E. Participants

We collected data over one week by contacting 250 volunteers through social networks. As a reminder, these 250 users were completely different from the 3,158 users of the training set. Moreover, none of them were part of a course on recommender systems, so as to avoid any bias.

We split them into 5 groups of 50 persons (G1 to G5). These 250 volunteers were 114 women and 136 men; 205 of them were French, the 45 others being from different countries over the world (Canada, Syria, Belgium, Roumania, Ivory Coast, Tunisia, Great Britain, Mexico, China, Lebanon, Algeria and Switzerland). There were 152 students, 62 senior executives, 25 employees, 5 retired persons, 4 self-employed people, 1 worker and 1 artisan. 4 of them were minors, 146 were between 18 and 24 years old, 65 were between 25 and 39 years old, 31 were between 40 and 59 years old, and 4 persons were more than 60 years old. Everybody, except one person, claimed going to theaters at least occasionally. They all were interested about movies. In order to motivate participants to diligently rate movies, they were told that there would be a lottery, that would reward 20 participants with a DVD in accordance with their preferences expressed within the frame of this study.

## IV. Results

### A. Measure of performance of algorithms

Before analyzing users' data, we measured the diversity level provided by each of the 5 algorithms in Step 2 (see Figure 1). As we used the average Intra-List Similarity measure from page 4 to page 10, the lower the similarity is, the more diverse the algorithm is. Let us remind that the 3 first pages were manually selected to initialize users' profiles. Thus, recommendations started at page 4. As expected, algorithms based on collaborative filtering (CF, CFRD, CFFD) provide much more diversity that CBF. Our diversity-based algorithms (CFRD and CFFD) are more diverse than the classical CF algorithm.

It is also not surprising that the POP algorithm provides a high level of diversity, since movies are randomly selected among the most popular. However, the POP algorithm does not provide any personalization since it does not rely on the active user' preferences. Thus, there is an important risk that users have a low confidence in these recommendations and/or do not find them relevant. The POP algorithm is only used as a baseline in our experiment.

At last, let us notice that the ILS measure decreases over time for the CBF algorithm, while it remains quite stable for the other algorithms. This is due to the small size of our movie corpus. Indeed, the CBF algorithm first recommends the movies that are the most similar as regards attributes with those that have been liked by the active user. Once it has recommended all the highly similar movies (movies that are at the same time from the same saga, with the same director, the same actors, the same popularity, and so on), it necessarily increases diversity by proposing movies that only have a few attributes in common.

### B. Validation of hypotheses

To validate our hypotheses, we analyzed results from the post-questionnaire (step 4). First, we converted answers into numerical values (from "Strongly disagree = 1" to "Strongly agree = 7"). Second, we computed answers' means for each of the group from G1 to G5 (see Table II).

**Validation of H1.** Groups G3 to G5, whose members used algorithms based on collaborative filtering in Step 2 (CF, CFRD, CFFD), found recommendations from the 5 algorithms more diverse in Step 3 than the other groups (see column "Diversity" in Tab II). We used a Student t-test to confirm the statistical significance of this result (p=0.05 between G2 and G3, p=0.07 between G2 and G4). Moreover, only 36 users from group G2 (CBF) found that the list of movies to be rated in Step 2 were diverse, against 45 to 47 users among a total of

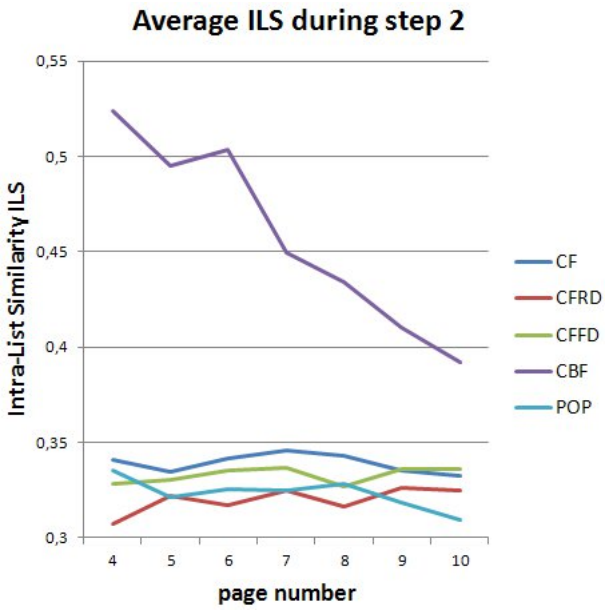


Fig. 1: Intra-List Similarities for each page in Step 2.

TABLE II: RESULTS FROM THE POST-QUESTIONNAIRE, AND MEANS OF RATINGS IN STEP 2

| Group number (algo. step 2) | Step 3 (all algorithms together) Diversity | Relevance | confidence | Mean in ratings |
|---|---|---|---|---|
| G1 (POP) | 4.64 | 3.94 | 4.98 | 3.49 |
| G2 (CBF) | 4.44 | 3.26 | 5.34 | 3.55 |
| G3 (CF) | **5** | 4.04 | 5.32 | 3.79 |
| G4 (CFRD) | **4.96** | 4.1 | **5.38** | 3.61 |
| G5 (CFFD) | 4.88 | **4.45** | 5.30 | 3.60 |

50 users for the other groups. Users are consequently capable of perceiving diversity within the recommendation set, even in the cases recommendations are made implicit (Step 2), which validates our hypothesis H1.

**Validation of H2.** The average ratings of collaborative filtering (CF) and diversity-based filtering (CFRD, CFFD) in Step 2 (column on the right in Tab II) are higher than those of CBF. This seems to confirm that diversity-driven algorithms (CF, CFRD, CFFD) improve users' satisfaction in comparison with other algorithms when recommendations are made in an implicit way. Nevertheless, the difference of satisfaction between these 3 variants of collaborative filtering (CF, CFRD, CFFD) remains marginal, and more particularly between CFRD and CFFD. Thus, we hypothesize that the degree of diversity does not have any impact on users' satisfaction, while a minimal threshold is reached. The latter will have to be clarified through another study where the degree of diversity will vary more finely, and on a greater number of recommendations.

**Validation of H3.** If diversity seems to improve satisfaction during the phase of implicit recommendation (Step 2), results are much more contrasted in Step 3 where users have been warned that the list of movies are recommended according to their explicit preferences. As shown in Tab III, we computed the number of times that each algorithm has been ranked first in Step 3, that is to say perceived by the user as the best TV channel. All groups together, we notice that CBF algorithm got the highest number of votes. This confirms hypothesis H3 according to which content-based filtering arises a higher level of user confidence (see the last line of Tab III). The comments provided by volunteers at the end of the study provides a piece of explanation: thanks to similarities of attributes, it is much easier for users to understand the link between preferences made explicit and recommendations from CBF, in comparison with other algorithms. As a consequence, each user can easily imagine an implicit explanation for a given recommendation (for example, the active user has highly rated the movie "The Matrix", which probably explains why the system recommends him/her the movie "The Matrix Reloaded").

TABLE III: NUMBER OF VOTES FOR EACH TV PROGRAM IN STEP 3

| Group Number | Algorithm chosen at Step 3 POP | CBF | CF | CFRD | CFFD |
|---|---|---|---|---|---|
| G1 (POP) | 14 | **22** | 7 | 3 | 4 |
| G2 (CBF) | 9 | **29** | 7 | 5 | **0** |
| G3 (CF) | 7 | **17** | 16 | 6 | 4 |
| G4 (CFRD) | 9 | **15** | 6 | 12 | 7 |
| G5 (CFFD) | **14** | 10 | 8 | 5 | 12 |
| confidence (all users together) | 4.98 | **5.34** | 5.32 | **5.34** | 5.32 |

On the other hand, according to the column entitled "Relevance" in Tab II, groups G4 and G5 – assigned to our diversity-driven algorithms in Step 2 (CFRD and CFFD) – found recommendations in Step 3 more relevant (all algorithms together) with more than one point of difference in comparison with group G2 assigned to content-based filtering. This result is statistically significant with a 99% level of confidence ($p = 0.004$ between G2 and G4, and $p = 4.27e-05$ between G2 and G5). Providing a more diverse set of items during Step 2 (CFRD) has also improved the overall degree of confidence of users within recommendations, even if the CBF algorithm got the highest number of votes in Step 3. As a consequence, whatever the recommendation algorithm used, the system has to make sure that the active user's preference model contains items diverse enough to provide better recommendations. This conclusion constitutes an unexpected influence of diversity, which will lead us to further investigate items that have to be rated during the cold-start phase.

## V. Conclusion and Perspectives

This work constitutes an explorative study of the role and impact of diversity within recommender systems. It highlighted the necessity to build preference models containing items various enough to ensure a good level of relevance and confidence of recommendations. Moreover, we proved that diversity is perceived by users and improve users' satisfaction. Nevertheless, diversity in the recommendation set can require additional explanations to users who may not see the link between their preferences made explicit and the items recommended by the system. In summary, diversity is a complex dimension, which is good for users, if it is used at the right time and in the appropriate manner. Following these conclusions, a perspective will consist in studying means to guarantee an adequate level of diversity during the cold-start phase.

## REFERENCES


[1] Netflix prize, http://www.netflixprize.com/, 2009.

[2] Y. Koren, R. M. Bell, and C. Volinsky, Matrix factorization techniques for recommender systems, IEEE Computer, vol. 42, no. 8, pp. 30–37, 2009.

[3] J. Sill, G. Takacs, L. Mackey, and D. Lin, Feature-weighted linear stacking, Cornell University, Netflix Prize Report, 2009.

[4] P. Sawers, Remember netflix's $1m algorithm contest? well, here's why it didn't use the winning entry, http://thenextweb.com/media/2012/04/13/remember-netflixs-1m-algorithm-contest-well-heres-why-it-didnt-use-the-winning-entry/, 2012.

[5] S. Castagnos, N. Jones, and P. Pu, Eye-tracking product recommenders' usage, in Proceedings of RecSys'10, Barcelona, pp. 29–36, 2010.

[6] N. Jones, User perceived qualities and acceptance of recommender systems, PhD Thesis, Ecole Polytechnique Fédérale De Lausanne, 2010.

[7] L. McGinty and B. Smyth, On the role of diversity in conversational recommender systems, in ICCBR'03, pp. 276–290, 2003.

[8] S. Castagnos, N. Jones, and P. Pu, Recommenders' influence on buyers' decision process, in In proc. of RecSys'09, New York, pp. 361–364, 2009.

[9] B. Smyth and P. McClave, Similarity vs. diversity, in Proceedings of the 4th International Conference on Case-Based Reasoning, Vancouver, pp. 347–361, 2001.

[10] Charles L.A. Clarke et al., Novelty and diversity in information retrieval evaluation, Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 659-666, 2008.

[11] F. Radlinski, P. N. Bennett, B. Carterette, and T. Joachims, Redundancy, diversity and interdependent document relevance, SIGIR Forum, vol. 43, no. 2, pp. 46–52, 2009.

[12] G. Adomavicius and Y. Kwon, Improving aggregate recommendation diversity using ranking-based techniques, IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 5, pp. 896–911, 2012.

[13] M. Zhang and N. Hurley, Avoiding monotony: Improving the diversity of recommendation lists, in Proceedings of RecSys'08, Lausanne, pp. 123–130, 2008.

[14] N. K. Lathia, Evaluating collaborative filtering over time, PhD Thesis, University College London, 2010.

[15] G. Häubl and K. Murray, Preference construction and persistence in digital marketplaces: The role of electronic recommendation agents, Journal of Consumer Psychology, vol. 13, no. 1, pp. 75–91, 2003.

[16] S. M. McNee, J. Riedl, and J. A. Konstan, Being accurate is not enough: how accuracy metrics have hurt recommender systems, in CHI '06: CHI '06 extended abstracts on Human factors in computing systems. Montréal: ACM, pp. 1097–1101, 2006.

[17] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl, Evaluating collaborative filtering recommender systems, ACM Transactions on Information Systems, vol. 22, pp. 5–53, 2004.

[18] Ana Beln Barragáns-Martínez et al., A hybrid content-based and item-based collaborative filtering approach to recommend TV programs enhanced with singular value decomposition, Journal of Information Sciences, Elsevier, vol. 180, pp. 4290-4311, 2010.

[19] C. Yu, L. V. Lakshmanan, and S. Amer-Yahia, Recommendation diversification using explanations, in Proceedings of the 2009 IEEE International Conference on Data Engineering (ICDE'09), pp. 1299–1302, 2009.

[20] M. Ge, F. Gedikli, and D. Jannach, Placing high-diversity items in top-n recommendation lists, in Workshop on Intelligent Techniques for Web Personalization and Recommender Systems (ITWP'11), Barcelona, Spain, pp. 65–68, 2011.

[21] F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, Recommender Systems Handbook. Springer, 2011.

[22] K. Bradley and B. Smyth, Improving recommendation diversity, in Irish Conference on Artificial Intelligence and Cognitive Science, Dublin, pp. 85–94, 2001.

[23] D. McSherry, Diversity-conscious retrieval, in Proceedings of EC-CBR'02, London, pp. 219–233, 2002.

[24] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong, Diversifying search results, in Proceedings of WSDM'09, Barcelona, pp. 5–14, 2009.

[25] Shengxian Wan et al., ICTNET at Web Track 2011 Diversity Track, Text REtrieval Conference (TREC'11), 2011.

[26] C.-N. Ziegler, S. McNee, J. Konstan, and G. Lausen, Improving recommendation lists through topic diversification, in Proceedings of the 14th international conference on World Wide Web (WWW'05), pp. 22–32, 2005.

[27] A. Said, B. Kille, B. J. Jain, and S. Albayrak, Increasing diversity through furthest neighbor-based recommendation, in Proceedings of the WSDM'12 Workshop on Diversity in Document Retrieval, Seattle, USA, 2012.

[28] M. Willemsen, B. Knijnenburg, M. Graus, L. Velter-Bremmers, and K. Fu, Using latent features diversification to reduce choice difficulty in recommendation lists, in RecSys'11 Workshop on Human Decision Making in Recommender Systems, Chicago, IL, pp. 14–20, 2011.

[29] B. P. Knijnenburg, M. C. Willemsen, Z. Gantner, H. Soncu, and C. Newell, Explaining the user experience of recommender systems, User Modeling and User-Adapted Interaction, vol. 22, no. 4-5, 2012.

[30] S. Castagnos, Website dedicated to a diversity-oriented experiment within recommender systems, http://www.movit.tv/tut5/index.php, 2013.

[31] Allociné website, http://www.allocine.fr/, 2013.

[32] V. Schickel and B. Faltings, Using an ontological a-priori score to infer user's preferences, in Workshop on Recommender Systems, in Conjunction with the 17th European Conference on Artificial Intelligence (ECAI 2006), pp. 102–106, August 2006.

[33] Imdb website, http://www.imdb.com/, 2013.

[34] B. Sarwar, G. Karypis, J. Konstan, and J. Reidl, Item-based collaborative filtering recommendation algorithms, in *World Wide Web*, pp. 285–295, 2001.