Literature Review

Jayachithra Kumar

December 3, 2017

Contents

1 Literature review **1**
1.1 Breaking the Filter-bubble . . . 1
1.2 Offline Evaluation (To do) . . . 1
1.2.1 Diversity Metrics . . . 1
1.2.2 Novelty metrics . . . 1
1.2.3 Serendipity Metrics . . . 2
1.2.4 Coverage metrics . . . 2
1.3 User-centric Evaluation . . . 3
1.3.1 Multi-criteria online studies . . . 3
1.3.2 Targeted online studies . . . 5
1.3.3 Discussion . . . 8
1.4 Comparison . . . 8
1.5 Conclusion . . . 8

Bibliography **9**

Chapter 1

Literature review

Article selection procedure

1.1 Breaking the Filter-bubble

1.2 Offline Evaluation (To do)

1.2.1 Diversity Metrics

Discuss definitions of diversity,different diversity metrics (they usually differ in the distance metrics used) and types of diversity (for eg. individual and aggregate diversity from Ge et al. (2011) [16]. Different metrics are:

1. Ziegler et al. (2005) [3] introduced topic diversification algorithm and intra-list similarity measure
2. Vargas and Castells (2011) [28] used Jaccard similarity - like metric
3. Ekstrand et al. (2014) [7] used intra-list similarity with cosine between tag genome vectors as itemwise similarity measure
4. Yu et al. (2009) [30] item distances are calculated based on neighborhood

Accuracy-diversity tradeoffs : "Solving the apparent diversity-accuracy dilemma of recommender systems" by Zhou et al. 2010, "Improving Aggregate Recommendation Diversity Using Ranking-Based Techniques" by Adamovicious et al. 2012.
Novelty and diversity - [28]

TABLE 1.1: Diversity metrics (to do)

Authors	individual diversity	aggregate diversity	accuracy-tradeoffs
Ziegler et al. (2005) [3]	X		
Vargas and Castells (2011) [28]			
Ekstrand et al. (2014) [7]	X		
Yu et al. (2009) [30]	X		
Zhou et al. (2011)		X	

1.2.2 Novelty metrics

Definition and methods to implement novelty

1. Baeza-Yates et al. (1999) gave the first discussion on novelty [10]
2. Zhang et al. (2002) [31] defined novelty based on distance between documents preciously seen by the user
3. distance-based novelty metrics - Yang et al. (2005) [29] and Nakatsuji et al. (2010) [5]
4. In [28], rank and relevance are taken into consideration
5. Novel items are unknown and not surprising to the user - [17] ; novel items also include known but forgotten items - [21]
6. increasing novelty by promoting "long - tail " items - [19], [23] and [8]

1.2.3 Serendipity Metrics

Definition of serendipity

1. Serendipitous items are both "attractive" and "surprising" to the users - Ge et al. (2010) [15]
2. SRDP measure relates the unexpectedness of an item with the item's usefulness - [15]
3. Measuring serendipity from *unexpectedness* measure - Murakami et al. (2007) [22]
4. Serendipity is a combination of *unexpectedness* and *item relevance* - Ge et al. (2010) [15]
5. An alternate measure for *unexpectedness* where *unexpectedness* is measured as a positive unbounded function of the distance of the item from a set of expected items - Adamopoulos (2011) [1]
6. *unserendipity* measure is used in Auralist [9]

1.2.4 Coverage metrics

Two types of coverage in literature - "user coverage" (extend to which user's are covered by the system) and "item coverage" (extend to which items are covered by the system).

1. Two types of item coverage - "prediction coverage" and "catalog coverage" proposed by Herlocker et al. (2004) [17]
2. Fleder et al. (2009) [14] proposed using Gini coefficient to measure the distribution of recommendation across all users
3. Shani et al. (2011) [26] used Shannon entropy to measure the distribution of recommendations across users
4. Relation between novelty and coverage - Jannach et al. (2013) [4]
5. Trade-off between coverage and serendipity - Ge et al. (2010) [15]

1.3 User-centric Evaluation

In order to better understand the performance of recommender systems from user's perspective, recent literature focuses on user-centric implementation and evaluation of recommender systems. One of the earliest applications of user-centric evaluations of recommender system focused on evaluating user's trust in recommendation interfaces and recommender agents [24]. User-centric evaluations are particularly important to measure beyond-accuracy objectives (i.e., diversity, novelty, coverage and serendipity) because without the help of end-user evaluations, it is not clear that an item that was identified to be serendipitous by the recommender algorithm is indeed perceived as the same by end-users. Furthermore, user-centric approaches help in validating the results of offline metrics and thereby informing researchers about different situations when the results of offline metrics can be trusted and when the results should be treated with skepticism. However, despite the obvious need for user-centric evaluations of beyond-accuracy objectives in recommender systems, most of the systems rely on just offline metrics and only a few works in this area actually include user-centric evaluations. The reason behind this could be justified by the high cost of conducting such user-centric evaluations, especially in designing the experiment - like setting the right questionnaires, finding appropriate number of participants etc. Moreover, user centric evaluations intend to measure subjective system aspects and hence careful design considerations - like choosing the order of displaying algorithms, measuring user expertise etc., - need to be made in order to avoid biased results. This section provides an overview of the limited number of works that rely on subjective evaluation of the beyond accuracy aspects of recommender systems.
Typically, all user-centric evaluation metrics use one of the two models - between-subjects design or within-subjects design. The main difference between both these models lie in the number of independent variables used in the experiment. In between-subjects design, for example, users are provided with one algorithm at a time whereas in within-subjects design, users compare multiple algorithms. In general, between-subjects design provide a more realistic view of the use of recommender system whereas within-subjects design can be used to evaluate and compare multiple algorithms. In addition to these two categories, Kaminskas et al. [2016] [20] provided a more specific classification where they categorize the research works based on the beyond accuracy aspects measured in the evaluation. Accordingly, the first category contains all research studies where the *relationships* between different user perceived recommendation qualities are measured, and the second category contains works that study the effect of specific algorithms or user-interface level adaptations on *specific* beyond-accuracy objectives. This classification is more particular for beyond-accuracy objectives and more relevant to the type of study that I intend to perform and hence, I choose to adopt this model of classification for rest of the discussion. For ease of representation, the categories are re-named as 'Multi-criteria online studies' and 'Targeted online studies' respectively without affecting the meaning of the categories.

1.3.1 Multi-criteria online studies

One of the first works based on multi-criteria user studies was performed by Pu et al. [2011] in their paper "A user-centric evaluation framework for recommender systems" [25]. They provide an extensive evaluation framework called ResQue which aimed at explaining how the perceived quality of the recommendation influences

user's beliefs, attitude towards the system and satisfaction with the system, and how these factors indeed influence user's behavioral intentions. They measure the perceived quality of the system with the help of two beyond accuracy objectives - novelty and diversity. Due to the meticulous distinction between novelty and serendipity, the latter was ignored from the study as it might confuse users. The framework consisted of an extensive set of 31 questions grouped into 15 categories. The experiment was performed with 239 participants of diverse nationalities where users were first asked to select a product of their choice from an online site, and then fill out the answers to the evaluation questionnaire from the framework. The results showed that the perceived usefulness of the system was considerably influenced by perceived novelty and moderately by the perceived diversity.

Knijnenburg et al. [2012] proposed an extensive framework for evaluating the user experience of recommender systems from objective system aspects by using subjective system aspects as mediators [2]. The framework is based on a set of six structurally related concepts namely objective system aspects (such as recommendation algorithm and presentation), subjective system aspects (such as perceived quality and diversity), perceived experience (i.e., attitude), interaction (i.e., behavior), situational characteristics (such as privacy concerns) and personal characteristics (such as trust). Several experiments were conducted to study the effects of one or more of these variables in the framework. User experience is measured from concepts such as perceived accuracy, diversity, system effectiveness, choice difficulty etc., by defining a chain of effects to draw relationships between these aspects. The results show inconsistent relation between actual and perceived diversity between the three algorithms used in the study (ie., k-nearest neighbor, matrix factorization and generally most popular algorithm). For example, the perceived diversity of k-NN algorithm with no actual diversity is higher than the perceived diversity of the same algorithm with a little actual diversity. However this condition does not hold for other two algorithms. Similarly an increase in perceived diversity tend to increase the perceived accuracy of the system thereby increasing the overall user experience. Finally, in case of "generally most popular" algorithm, diversification of the algorithm is shown to be as effective as replacing it with a recommendation algorithm.

In "User perception of differences in recommender algorithms" [7], Ekstrand et al. compared three collaborative filtering algorithms - item-item CF, user-user CF and singular value decomposition (SVD) CF -, on five dimensions - novelty, diversity, accuracy, satisfaction and personalization. The framework of Knijnenburg et al., [2] was used to model the evaluation and MovieLens user community was recruited as participants. A within-subjects study was conducted and each user was assigned to two out of the three algorithms. In the first step, users were provided with two lists of movies with 10 recommendations side-by-side on the same interface, and they were asked to choose their most preferred list based on first impression. Secondly, users were asked to answer a set of 22 questions about various aspects of the list, and finally users were asked to choose their most preferred algorithm. The results show that the perceived values of diversity and novelty correlate with the measured values, and that, diversity has a positive influence on user's choice of the system and novelty has a significant negative impact on user's satisfaction. As a result, user-user CF algorithm, which has highest novelty among the three algorithms, was least preferred compared to the other two algorithms.

Finally, Fazeli et al. [2017] [13] in their work, try to find the relation between user-satisfaction of a recommender system measured using online evaluation and the accuracy of the system measured offline. User satisfaction is evaluated based on five quality metrics such as perceived usefulness, accuracy, novelty, diversity

and serendipity which was taken from the ResQue framework. The experiment used a between-subjects design and involved users of Open Discovery Space (ODS) which is an eLearning environment. Three best algorithms were chosen from the offline evaluation and these were used for online evaluation. These algorithms are a memory-based CF (User KNN), a graph-based CF and a model-based CF(matrix factorization). Each user evaluated recommendations from a randomly assigned algorithm and each algorithm was assigned to 20 users. The questionnaire consisted of a set of six questions that measure the five quality metrics. Results reveal that, even though traditional evaluation suggests User KNN as the best algorithm, users were satisfied with the accuracy of all the algorithms regardless of the choice of the algorithm. The authors suggest that there is no point in finding the most accurate algorithm if users do not recognize the differences between recommender systems.

TABLE 1.2: Coverage of beyond-accuracy objectives in multi-criteria online studies

Authors	Novelty	Diversity	Serendipity	Coverage
Pu et al. [25]	X	X		
Knijnenburg et al. [2]		X		
Ekstrand et al. [7]	X	X		
Fazeli et al. [2017] [13]	X	X	X	

Using the framework suggested in Knijnenburg et al [2] as baseline, comparisons between the above three works were made (table to be added). The impact of personal and situational characteristics on user satisfaction remain one of the understudied concepts in these studies. However, from the results of experiments in [2], it is clearly evident that these characteristics do tend to influence the overall experience of the user to a great extend. For example, users with higher domain expertise tend to have higher perceived diversity compared to users with lower expertise. Furthermore, from table 1.2 it is evident that, out of the four beyond accuracy objectives, only novelty and diversity are mostly addressed in the studies. Serendipity was ignored in Pu et. al [25] in order to avoid confusing users. One reason for not measuring the concept of coverage could be because it has multiple dimensions and that it is defined at system level and not user level. Finally on comparing the results of the four studies we notice that while the study by Pu et al., [25] implies an increase in novelty increases user satisfaction, the study by Ekstrand et al. [7] seems to contradict this result. One reason for this could be attributed to the differences in question formation. While in [25], the users were directly asked to answer *"The items in the list are novel"*, in [7], novelty was measured from indirect questions like *"has more movies that you do not expect?"* etc. Furthermore, the nature of the question *"has more pleasantly surprising movies?"* in [7] suggests that it correlates to 'serendipity' aspect more than 'novelty' of recommendation and this might have introduced a certain bias in the result.

1.3.2 Targeted online studies

Recent literature provides several targeted user studies which were conducted in order to evaluate a specific beyond-accuracy criteria. In general, online studies evaluate one of the two methods of implementations of beyond-accuracy objectives - *algorithm-level implementation*, where beyond accuracy objectives are included by

tuning the underlying recommendation algorithm, and, *interface-level implementation*, where the user-interface is modified in order to make users to perceive and/or consume novel, diverse or serendipitous elements. User-interface level implementations consist of either *explanation interfaces* [27] - where explanations are provided for unusual recommendations-, or *organization interfaces* [18] - where results are reorganized and grouped to improve classification.

One of the earliest works on user-evaluation of recommender systems were done by Ziegler et al. [2005] [3] where a topic-diversification algorithm was proposed in order to increase the diversity of recommendation lists. Evaluation of the algorithm was done both offline - using the intra-list similarity metric and online - with more than 2,100 users. The experiment was performed with data from BookCrossing.com and Amazon.com and the effects of diversity was measured on two collaborative filtering (CF) algorithms - item-item CF and user-user CF. A between-subject study was conducted where users were first displayed with a list of 10 recommendations from one of the two algorithms and asked to rate the items in the list after which they were presented with a set of questions that measure user's perceived diversity and satisfaction with the recommendations. The results showed that an increase in diversification tend to increase overall user satisfaction, which was notably significant in case of item-item CF algorithm.

The experiment by Celma in 2009 [12] was aimed at measuring the Novelty of the recommendation systems by explicitly asking users if they were familiar with the item or not. A within-subjects study was conducted in order to compare the novelty in three algorithms - CF, content-based audio similarity (CB) and a hybrid approach that combines CB with Allmusic.com's "human expert information". The study was performed on last.fm users and consisted of several rounds. In each round, participants were asked to rate songs from a list of 10 recommended songs and the rating was collected based on familiarity - whether the user knows the song - and quality - whether the user likes the song. Novelty is considered to be the inverse of familiarity and the results show that the list with higher novelty was considered to be of lower quality compared to the list with higher familiarity. Adding meta-data and explanation for why a particular song was recommended was proposed as a possible solution to improve user's acceptance of novel recommendations.

Willemsen et al [2011] [6] conducted a within-subject user study to prove that increasing the diversity of a recommendation list helps in effectively decreasing the choice-overload problem. The experiment was conducted on a movie recommendation system that uses matrix factorization algorithm. The algorithm was tuned to provide three different levels of diversity and user study was conducted to evaluate how different levels of diversification affected the perceived diversity of the list. The study consisted of 97 participants and each participant was asked to rate three different lists of varying diversity, followed by a set of questions to measure perceived diversity, perceived attractiveness, expertise, choice difficulty and trade-off difficulty. The evaluation framework was based on [2] and the results showed that increasing the diversity of the list tend to increase the perceived diversity and decrease the choice and trade-off difficulty.

In the same year, Hu et al. [18] studied how an organization interface compares to a standard list interface in terms of perceived usefulness and diversity. In an organization interface, the recommendations are categorized based on some common trade-off properties (e.g. products which are "cheaper but heavier" than the chosen product are grouped together). The experiment was conducted on a commercial perfume website and a total of 20 participants were recruited for the study. A within-subjects study was conducted and the evaluation framework was based on ResQue

from [25]. All participants used both the interfaces assigned to them in random order and answered a set of questions about their preferences. The questions were based on user's general preference, usefulness, informativeness, etc and they intended to measure two main aspects - *categorical diversity* (difference among categories) and *item-item diversity* (difference among items). The results showed that users perceived categorical diversity much stronger than item-item diversity and that they perceived organization interface to be much more useful than list interface.

Ge et al., in their paper "Placing high diversity items in top-n recommendation list" [16], attempt to find the efficient placement of diverse items in the recommendation list and the extend to which diversity influences perceived quality of the recommendation system. Static lists of movies were created for three genre with three different placements of diverse items. All diverse items were placed together in a single block - at the end of the list for action genre and in the middle of the list for romance genre. Diverse items were dispersed into different positions for comedy genre. A within-subjects pilot study was conducted with 10 users and each user was provided with all the three lists, with the genre difference as the determinant of diversity. Users were asked to rate movies and answer four questions related to user's satisfaction, diversity and surprise. Results showed that diverse items in the list aroused user's interest and attention and that discovery of diverse items were much more effective when they were arranged in a block than when they were distributed across the list. The results also showed that users were more interested to read additional information about diverse items, thereby stressing the importance of providing explanation interfaces in recommender systems.

Zhang et al. [9] conducted a user-study on their serendipitous framework called *Auralist* in order to study the impact of serendipity, novelty and diversity on user satisfaction. They develop three hybrid versions of their basic music recommender system. The first one called *Community-Aware Auralist* aims to provide diversity; the second algorithm called *Bubble-Aware Auralist* recommends artists outside user's music-bubble; and finally the third one *Full Auralist* is a hybrid of first and second algorithm. User study involved 21 participants and it was aimed at comparing the perceived serendipity, novelty, enjoyment and overall qualitative satisfaction between *Basic* and *Full Auralist*. Each user was presented with 20 recommendation from both the recommendation algorithms in random order and for each algorithm questions were asked to measure user's serendipity, novelty, enjoyment and usefulness. Results showed that users found *Full Auralist* to be more useful than *Basic Auralist* even though there is a slight compromise in accuracy in the former algorithm. Serendipity and novelty is seen as a positive contributor to user's satisfaction.

Finally, Castagnos et al. [2013] [11] measured the impact of diversity on user-satisfaction in three different algorithms - CF, CB and Popularity-based filtering (POP). The experiment was conducted in movie domain and the study involved four steps. Each participant was first asked to fill out a questionnaire on demographics and expertise. Users were then asked to rate movies provided by one of the three algorithms assigned to him/her and in the third step they were provided with three recommendation lists from the three algorithms and were asked to order the list based on their preferences. Finally, a post-questionnaire was presented to measure the perceived relevance, diversity and confidence level of the recommendation. The results correlate with the other studies showing that diversity has a positive influence on user's satisfaction. However, too much diversity tend to confuse users and have a negative impact if proper explanation is not provided.

On comparing the works on targeted online studies (table 1.3), two interesting observations can be made. Firstly, it is clear that serendipity and coverage remain

TABLE 1.3: Coverage of beyond-accuracy objectives in targeted online studies

Authors	Novelty	Diversity	Serendipity	Coverage
Ziegler et al. (2005) [3]		X		
Celma (2009) [12]	X			
Willemsen et al. (2011) [6]		X		
Hu et al. (2011) [18]		X		
Ge et al. (2011) [16]		X		
Zhang et al. (2012) [9]	X		X	
Castagnos et al. (2013) [11]		X		

under-explored compared to diversity and novelty. However, on further analysis it is evident that the two main aspects of serendipity, i.e., "surprise" and "usefulness" are measured independently in several studies. For example, in Ge et al. (2011) [16], the authors measure how surprising the recommendations are to the users and in Hu et al. (2011) [18] and Zhang et al. (2012) [9], the authors measure the perceived usefulness of the recommender systems. Therefore by including questions that correlate "surprise" and "usefulness" of a list we could possibly measure the perceived serendipity of the recommendation system. Secondly, the results of four out of seven studies imply that providing explanations and meta-data for recommendation items play a crucial role in shaping user's perception of diversity. These results were obtained from experiments in three different domains - [11] and [16] in movie domain, [12] in music domain and [18] in e-commerce - and hence it is evident that explanation interfaces play a crucial role in increasing the perceived usefulness of recommender systems irrespective of the application domain.

1.3.3 Discussion

1.4 Comparison

1.5 Conclusion

Bibliography

[1] Panagiotis Adamopoulos and Alexander Tuzhilin. “On Unexpectedness in Recommender Systems: Or How to Expect the Unexpected”. In: (2011).

[2] Bart P. Knijnenburg et al. “Explaining the user experience of recommender systems”. In: (2012).

[3] Cai-Nicolas Ziegler et al. “Improving recommendation lists through topic diversification”. In: (2005).

[4] Dietmar Jannach et al. “What recommenders recommend–an analysis of accuracy, popularity, and sales diversity effects”. In: (2013).

[5] Makoto Nakatsuji et al. “Classical music for rock fans?: novel recommendations for expanding user interests”. In: (2010).

[6] Martijn C. Willemsen et al. “Using latent features diversification to reduce choice difficulty in recommendation lists”. In: (2011).

[7] Michael D. Ekstrand et al. “User perception of differences in recommender algorithms”. In: (2014).

[8] Tao Zhou et al. “Solving the apparent diversity-accuracy dilemma of recommender systems”. In: (2010).

[9] Yuan Cao Zhang et al. “Auralist: introducing serendipity into music recommendation”. In: (2012).

[10] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. “Modern information retrieval”. In: (1999).

[11] Armelle Brun Castagnos Sylvain and Anne Boyer. “When Diversity Is Needed... But Not Expected!” In: (2013).

[12] Òscar Celma Herrada. “Music recommendation and discovery in the long tail”. In: (2009).

[13] Soude et al Fazeli. “User-centric Evaluation of Recommender Systems in Social Learning Platforms: Accuracy is Just the Tip of the Iceberg”. In: (2017).

[14] Daniel Fleder and Kartik Hosanagar. “Blockbuster culture’s next rise or fall: The impact of recommender systems on sales diversity”. In: (2009).

[15] Carla Delgado-Battenfeld Ge Mouzhi and Dietmar Jannach. “Beyond accuracy: evaluating recommender systems by coverage and serendipity”. In: (2010).

[16] Fatih Gedikli Ge Mouzhi and Dietmar Jannach. “Placing high diversity items in Top-N Recommendtion Lists”. In: (2011).

[17] Jonathan L. et al. Herlocker. “Evaluating collaborative filtering recommender systems”. In: (2004).

[18] Rong Hu and Pearl Pu. “Enhancing recommendation diversity with organization interfaces”. In: (2011).

[19] Masayuki et al Ishikawa. “Long tail recommender utilizing information diffusion theory”. In: (2008).

[20] Derek Bridge Kaminskas Marius. “Diversity, Serendipity, Novelty, and Coverage: A Survey and Empirical Analysis of Beyond-Accuracy Objectives in Recommender Systems”. In: (2016).

[21] Komal et al. Kapoor. “I like to explore sometimes: Adapting to dynamic user novelty preferences”. In: (2015).

[22] Koichiro Mori Murakami Tomoko and Ryohei Orihara. “Metrics for evaluating the serendipity of recommendation lists”. In: (2007).

[23] Yoon-Joo Park and Alexander Tuzhilin. “The long tail of recommender systems and how to leverage it”. In: (2008).

[24] Li Chen Pu Pearl. “Trust building with explanation interfaces”. In: (2006).

[25] Li Chen Pu Pearl and Rong Hu. “A user-centric evaluation framework for recommender systems”. In: (2011).

[26] Guy Shani and Asela Gunawardana. “Evaluating recommendation systems.” In: (2011).

[27] Judith Masthoff Tintarev Nava. “Explaining recommendations: Design and evaluation”. In: (2015).

[28] Saúl Vargas and Pablo Castells. “Rank and relevance in novelty and diversity metrics for recommender systems”. In: (2011).

[29] Yan Yang and Jian Zhong Li. “Interest-based recommendation in digital library.” In: (2005).

[30] Laks VS Lakshmanan Yu Cong and Sihem Amer-Yahia. “Recommendation diversification using explanations”. In: (2009).

[31] Jamie Callan Zhang Yi and Thomas Minka. “Novelty and redundancy detection in adaptive filtering”. In: (2002).