SURVEY PAPER

A Survey on Recommendation Methods Beyond Accuracy

Jungkyu HAN[†a)], *Nonmember and* **Hayato YAMANA**[†b)], *Member*

SUMMARY In recommending to another individual an item that one loves, accuracy is important, however in most cases, focusing only on accuracy generates less satisfactory recommendations. Studies have repeatedly pointed out that aspects that go beyond accuracy—such as the diversity and novelty of the recommended items—are as important as accuracy in making a satisfactory recommendation. Despite their importance, there is no global consensus about definitions and evaluations regarding beyond-accuracy aspects, as such aspects closely relate to the subjective sensibility of user satisfaction. In addition, devising algorithms for this purpose is difficult, because algorithms concurrently pursue the aspects in trade-off relation (i.e., accuracy vs. novelty). In the aforementioned situation, for researchers initiating a study in this domain, it is important to obtain a systematically integrated view of the domain. This paper reports the results of a survey of about 70 studies published over the last 15 years, each of which addresses recommendations that consider beyond-accuracy aspects. From this survey, we identify diversity, novelty, and coverage as important aspects in achieving serendipity and popularity unbiasedness—factors that are important to user satisfaction and business profits, respectively. The five major groups of algorithms that tackle the beyond-accuracy aspects are multi-objective, modified collaborative filtering (CF), clustering, graph, and hybrid; we then classify and describe algorithms as per this typology. The off-line evaluation metrics and user studies carried out by the studies are also described. Based on the survey results, we assert that there is a lot of room for research in the domain. Especially, personalization and generalization are considered important issues that should be addressed in future research (e.g., automatic per-user-trade-off among the aspects, and properly establishing beyond-accuracy aspects for various types of applications or algorithms).

key words: *recommendation, beyond-accuracy, diversity, novelty, long-tail*

1. Introduction

Accuracy with respect to the ability to recommend items loved by a user is important to making a satisfactory recommendation. In most cases, however, only pursuing accurate recommendations therefore ignoring all other aspects related to user satisfaction makes delivering satisfactory recommendations to users difficult. For instance, recommending a new Star Wars movie to a Star Wars fan or filling their recommendation list with Star Wars series may be accurate but less satisfactory, as they have many ways of knowing about the movies. In such a case where a user expects "the movies loved by me, but *difficult to be found* by myself" from recommendation systems, the recommendation systems cannot fulfill his/her satisfaction. These kinds of recommendations also act negatively in business. With what is known as "long-tail," a substantial proportion of the profits made by on-line retailers comes from a large number of items that do not receive the attention of most users. For example, 30% of Amazon.com's profits comes from purchases of less-popular niche products [5]. Accurate recommendations of well-known items cannot promote such items.

Most studies that focus on accuracy tend to suffer from the aforementioned problems in some way. It is well known that collaborative filtering (CF), one of the popular algorithms in the literature of recommendation systems, is accurate, but it recommends users the items similar to each other. Relatively recently, studies have repeatedly pointed out that beyond-accuracy aspects—such as the diversity and novelty of the recommended items—should be considered alongside accuracy [22], [45].

Despite the importance of beyond-accuracy aspects, there is no global consensus regarding definitions and evaluations of the aspects, as such aspects closely relate to the subjective sensibility of user satisfaction. In addition, devising algorithms is difficult, as algorithms should find an ideal balance between the aspects in trade-off relation (e.g., accuracy vs. novelty) to achieve maximum user satisfaction. Given such circumstances, for researchers initiating a study in this domain, obtaining a systematically integrated view of the domain is important. Nonetheless there has been no comprehensive review of recommendation methods that incorporate beyond-accuracy aspects. This paper reports the results of a survey of studies on recommendations adopted through the use of beyond-accuracy aspects. Popular beyond-accuracy aspects—including diversity and novelty—are identified, and the definition of each aspect is presented, based on the contents of the surveyed studies. Then, algorithms that consider beyond-accuracy aspects, as well as off-line evaluation metrics, are classified and described in terms of their purpose and methodologies. The user studies carried out by the surveyed studies are also described.

This paper is organized as follows. In Sect. 2, we describe the research questions for this survey, as well as the selection methodology used. The detailed results of the survey are reported in Sect. 3. We describe open questions and challenges to be studied in Sect. 4. Section 5 provides concluding remarks.

Manuscript received March 9, 2017.
Manuscript revised July 11, 2017.
Manuscript publicized August 23, 2017.

[†]The authors are with Waseda University, Tokyo, 159–8555 Japan.
a) E-mail: han.jungkyu@akane.waseda.jp
b) E-mail: yamana@waseda.jp

DOI: 10.1587/transinf.2017EDR0003

2. Research Questions and Surveyed Papers

In this section, we describe the research questions that prompted this survey, as well as the way in which we selected the papers to be surveyed.

2.1 Research Questions

Despite the importance of beyond-accuracy aspects, there exist relatively fewer studies on them than on accurate recommendations. The phenomenon mainly originated from (1) the strong relationship between beyond-accuracy aspects and subjective satisfaction of users, and (2) the trade-off relation between beyond-accuracy aspects and accuracy. Given the subjectivity involved, various definitions and interpretations about beyond-accuracy aspects have been proposed. The trade-off relation with accuracy makes it difficult to devise algorithms that simultaneously satisfy both accuracy and beyond-accuracy aspects to their maximum. As a consequence, most algorithms seek an acceptable trade-off between accuracy and beyond-accuracy aspects. Because user satisfaction is subjective perception, the trade-off makes it difficult to perform effective off-line evaluations, for in such circumstances—and in the absence of a real user—we do not have concrete criteria by which to assess the goodness of a given trade-off.

Our research questions were formulated so as to guide researchers who have an interest in this domain. Table 1 shows our three research questions. Because various opinions coexist, constructing a systematic view of the definitions from related studies can provide a thorough overview of the domain, and therefore help researchers in identifying problems clearly (RQ 1). In classifying the algorithms that consider beyond-accuracy aspects and understanding the ideas behind the algorithms, researchers can derive some insights into "What has been accomplished and what are the problems to be solved?" (RQ 2). Finally, we believe that securing a thorough knowledge of evaluation methodology will help researchers carry out sound evaluations of their proposed methods (RQ 3).

2.2 Selection of Surveyed Papers

To select the papers to be surveyed, we first built a seed paper set. The papers in the seed set were selected by undertaking a manual inspection of the publication lists of the four international conferences within the predefined period. The publication lists of SIGIR[†], KDD[††], and WWW[†††] were scanned for a recent five-year period (i.e., 2012–2016). RecSys[††††] was scanned for the whole of its 10-year history (i.e., 2007–2016), as RecSys is the conference that most closely relates to our research area. We selected papers that met both of the following criteria.

- **Criteria 1.** The title or abstract contains one of the following words: diversity, serendipity, novelty, popularity bias, long-tail and coverage. Or their synonyms: divers-e/-ification/-fing/-ified, heterogeneous, serendipitous, unexpectedness, novel, discovery, popularity, popularity-/selection-bias, over-specialization, short-head, concentrat-ion/-ed, and niche.

- **Criteria 2.** The abstract relates to recommendation.

In this survey, we focused on the beyond-accuracy aspects related to the "satisfactory recommended items." The beyond-accuracy aspects represented by the keywords in Criteria 1 are directly related to the user satisfaction with the recommended items. Besides of the aspects represented by the keywords, a variety of important beyond-accuracy aspects are discussed in the literature of recommendation systems. For instance, the amount of training data required for the recommendation with sufficient accuracy [29], comprehensibility of recommendation (explicable recommendation), and interfaces [69] are important aspects that should be considered. Although the other beyond accuracy-aspects such as comprehensibility and interfaces are important, we think that the items with high user satisfaction should be recommended first because recommending such items is a necessary condition for the user satisfaction improvement achieved by the aforementioned aspects. Therefore we selected the keywords in Criteria 1, and in this paper, the term "beyond-accuracy aspects" indicates the aspects represented by the keywords in Criteria 1.

Table 1 Research questions

Category	Main question	Sub-question	Section
Definitions	RQ 1. What are the definitions of beyond-accuracy aspects in recommendation system?	N/A	3.1
Algorithms	RQ 2. What kinds of approaches developed and what are the representative idea of each kind?	N/A	3.3
Evaluations	RQ 3. How beyond-accuracy aspect aware recommendation systems are evaluated?	RQ 3-1. Evaluation metrics in off-line evaluation	3.2
		RQ 3-2. User study	3.4

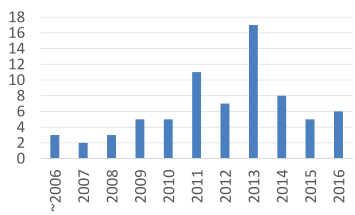


Fig. 1 The number of publications w.r.t. year

[†]International ACM SIGIR Conference on Research and Development in Information Retrieval
[††]SIGKDD Conference on Knowledge Discovery and Data Mining
[†††]International World Wide Web Conference
[††††]ACM Conference on Recommender Systems

Table 2 Surveyed paper list

Year	Authors [ref]	Objective	Alg.	App. domain	Acc.	Div.	Nov.	Cov.	Seren.	P-Unbias.	Off. / U-Std.
2001	Bradley and Smyth [11]	Algorithm	Multi Obj.	Job	✓	✓	-	-	-	-	✓/ -
2002	Bridge and Ferguson [12]	Algorithm	Hybrid	Case Retr.	✓	✓	-	-	-	-	✓/ -
2005	Ziegler *et al.* [96]	Algorithm	Multi Obj.	Books	✓	✓	-	-	-	-	✓/✓
2007	Murakami *et al.* [47]	Eval. metric	-	TV show	✓	-	✓	-	-	-	✓/ -
2007	Weng *et al.* [83]	Algorithm	Hybrid	-	-	-	-	-	-	-	- / -
2008	Park *et al.* [54]	Algorithm	Others	Movies, Books	✓	-	-	-	-	-	✓/ -
2008	Celma and Herrera [16]	Eval. metric	-	Music	-	-	✓	-	-	-	✓/✓
2008	Ishikawa *et al.* [33]	Algorithm	Others	Tech. articles	✓	-	✓	-	-	✓	✓/ -
2009	Zhang and Hurley [88]	Algorithm	Clustering	Movies	✓	✓	✓	-	-	✓	✓/ -
2009	Hijikata *et al.* [30]	Algorithm	Hybrid	Music	✓	-	✓	-	✓	✓	✓/✓
2009	Yu *et al.* [86]	Algorithm	Multi Obj.	Movies, Tags	✓	✓	-	-	-	-	✓/ -
2009	Zhang and Hurley [90]	Algorithm	Others	Movies	✓	✓	-	-	-	-	✓/ -
2009	Zhang and Hurley [89]	Eval. metric	-	Movies	✓	✓	✓	✓	-	✓	✓/ -
2010	Ge *et al.* [26]	Eval. metric	-	-	-	-	-	✓	✓	-	- / -
2010	Zhou *et al.* [94]	Algorithm	Hybrid	Movies, Music, ...	✓	-	✓	-	-	✓	✓/ -
2010	Zhang and Hurley [91]	Algorithm	Multi Obj.	Movie	✓	-	✓	-	-	✓	✓/ -
2010	Jambor and Wang [34]	Algorithm	Multi Obj.	Movie	✓	✓	✓	-	-	-	✓/ -
2010	Lathia [40]	Investigation	-	Movies	✓	✓	✓	-	-	-	✓/ -
2011	Hurley and Zhang [31]	Algorithm	Multi Obj.	Movie, Profile, ...	✓	✓	✓	-	-	-	✓/ -
2011	Boim *et al.* [10]	Algorithm	Clustering	Movies	✓	✓	-	-	-	-	✓/✓
2011	Li *et al.* [42]	Algorithm	Clustering	News articles	✓	✓	-	-	-	-	✓/✓
2011	Wartena and Wibbels [81]	Algorithm	Clustering	Books	✓	✓	-	-	-	-	✓/ -
2011	Steck [67]	Algorithm	Mod. of CF	Moive	✓	-	✓	✓	-	-	✓/✓
2011	Lee and Lee [41]	Algorithm	Mod. of CF	Music	✓	-	✓	-	-	-	- /✓
2011	Vargas and Castells [75]	Eval. metric	-	Movies, Music	✓	✓	✓	-	-	-	✓/ -
2011	Castells and Vargas [15]	Eval. metric	-	Movies	✓	✓	✓	-	-	-	✓/ -
2011	Oh *et al.* [51]	Algorithm	Multi Obj.	Movies	✓	✓	✓	✓	-	-	✓/ -
2011	Vargas *et al.* [76]	Algorithm	Multi Obj.	Movies	✓	✓	-	-	-	-	✓/ -
2011	Tong *et al.* [74]	Algorithm	Multi Obj.	Citation	✓	✓	-	-	-	-	✓/ -
2012	Li and Murata [43]	Algorithm	Clustering	Movies	✓	-	-	-	-	-	✓/ -
2012	Shi and Ali [65]	Algorithm	Mod. of CF	Moves, Apps.	✓	-	✓	-	-	-	✓/ -
2012	Yin *et al.* [85]	Algorithm	Graph	Movies, Books	✓	-	✓	✓	✓	✓	✓/✓
2012	Zhang *et al.* [92]	Algorithm	Hybrid	Music	✓	✓	✓	-	✓	-	✓/✓
2012	Adomavicius and Kwon [3]	Algorithm	Multi Obj.	Movies	✓	-	✓	✓	-	✓	✓/ -
2012	Ribeiro *et al.* [58]	Algorithm	Multi Obj.	Movies, Music	✓	✓	✓	-	-	-	✓/ -
2012	Shi *et al.* [66]	Algorithm	Others	Movies	✓	✓	-	-	-	-	✓/ -
2013	Zhao *et al.* [93]	Algorithm	Mod. of CF	Movies	✓	-	✓	✓	-	✓	✓/ -
2013	Niemann and Wolpers [49]	Algorithm	Mod. of CF	Movies	✓	-	✓	✓	-	✓	✓/ -
2013	Said *et al.* [61]	Algorithm	Mod. of CF	Movies	✓	-	✓	-	✓	-	✓/✓
2013	Qin and Zhu [56]	Algorithm	Multi Obj.	Movies	✓	✓	-	-	-	-	✓/ -
2013	Hurley [32]	Algorithm	Mod. of CF	Movies	✓	✓	-	-	-	-	✓/ -
2013	Shi [64]	Algorithm	Graph	Movies, Music	✓	-	✓	✓	-	-	✓ -
2013	Su *et al.* [68]	Algorithm	Mod. of CF	Movies	✓	✓	-	-	-	-	✓/ -
2013	Taramigkou *et al.* [71]	Algorithm	Graph	Music	✓	-	✓	-	✓	-	- /✓
2013	Mourão *et al.* [46]	Algorithm	Hybrid	Movies, Music	✓	✓	✓	-	-	-	✓/ -
2013	Belém *et al.* [8]	Algorithm	Multi Obj.	Tags(Mov., Mus.)	✓	✓	✓	-	-	-	✓/ -
2013	Belém *et al.* [7]	Algorithm	Multi Obj.	Tags(Mov., Mus.)	✓	✓	-	-	-	-	✓/ -
2013	Küçüktunç *et al.* [38]	Algorithm	Multi Obj.	Cit. or Soc. NW	✓	✓	-	-	-	-	✓/ -
2013	Vargas and Castells [77]	Algorithm	Multi Obj.	Movies, Music	✓	✓	-	-	-	-	✓/ -
2013	Abbar *et al.* [1]	Algorithm	Multi Obj.	News articles	✓	✓	-	-	-	-	✓/✓
2013	Servajean *et al.* [63]	Algorithm	Multi Obj.	Papers	✓	✓	-	-	-	-	✓/ -
2013	Kohli *et al.* [37]	Algorithm	Others	Movies, Jokes	✓	-	-	-	-	-	✓/ -
2013	Szpektor *et al.* [70]	Algorithm	Others	Questions	✓	-	-	-	-	-	✓/✓
2014	Adamopoulos and Tuzhilin [2]	Algorithm	Mod. of CF	Movies, Foods	✓	-	✓	✓	-	-	✓/ -
2014	Vargas and Castells [78]	Algorithm	Mod. of CF	Movies, Music	✓	-	✓	✓	-	✓	✓/ -
2014	Ekstrand *et al.* [21]	Investigation	-	Movies	✓	✓	✓	-	✓	-	- /✓
2014	Panniello *et al.* [52]	Investigation	-	Goods	✓	✓	-	-	-	-	✓/ -
2014	Nguyen *et al.* [48]	Investigation	-	Movies	✓	✓	-	-	-	-	✓/ -
2014	Cremonesi *et al.* [20]	Investigation	-	Hotels	✓	-	✓	-	✓	✓	- /✓
2014	Vargas *et al.* [79]	Algorithm	Multi Obj.	Movies	✓	✓	-	-	-	-	✓/ -
2014	Noia *et al.* [50]	Algorithm	Multi Obj.	Movies	✓	✓	-	-	-	-	✓/ -
2015	Ashkan *et al.* [6]	Algorithm	Mod. of CF	Movies	✓	✓	-	-	-	-	✓/✓
2015	Chatzicharalampous *et al.* [17]	Algorithm	Mod. of CF	Movies, Music	✓	✓	✓	✓	-	-	✓/ -
2015	Christoffel *et al.* [18]	Algorithm	Graph	Movies, Books	✓	-	✓	✓	-	✓	✓/ -
2015	Küçüktunç *et al.* [39]	Algorithm	Graph	Citations	✓	✓	-	-	-	-	✓/ -
2015	Kapoor *et al.* [36]	Algorithm	Other	Music	✓	-	✓	-	-	-	✓/ -
2016	Wu *et al.* [84]	Algorithm	Mod. of CF	Movies, Books, ...	✓	✓	-	-	-	-	✓/ -
2016	Tobìas *et al.* [73]	Investigation	-	Movies, Music	✓	✓	-	-	-	-	✓/ -
2016	Parambath *et al.* [53]	Algorithm	Multi Obj.	Movies	✓	✓	✓	✓	-	-	✓/ -
2016	Teo *et al.* [72]	Algorithm	Multi Obj.	Goods	✓	✓	-	-	-	-	- /✓
2016	Wasilewski and Hurley [82]	Algorithm	Multi Obj.	Movies	✓	✓	-	-	-	-	✓/ -
2016	Benouaret and Lenne [9]	Algorithm	Multi Obj.	POIs	✓	✓	-	-	-	-	✓/ -

After the seed set was constructed, we extended it by using the references of each paper within the set, while imposing the two aforementioned criteria. This extension was executed in two iterations. We then looked for duplicate results. For instance, if one paper was a journal version of a paper previously published by the same authors, we included the most recently published paper. If a paper only describes the authors' previous work, then we replaced it with the previous papers of the same authors. We also excluded problem-raising papers that accentuated the im-

portance of beyond-accuracy aspects based on real-world data [5] or simulation [22]. Ultimately, we selected 72 papers for review.

2.3 Summary

Figure 1 plots the number of selected publications in each year. The oldest paper among those selected for our survey was published in 2001 [11]. Table 2 summarizes the selected papers. The "Objective" column indicates the study objectives: most of the studies focused on devising recommendation *algorithms*. Other studies discussed *evaluation metrics* for beyond-accuracy aspects, or *investigated* the performance of conventional algorithms from the perspective of beyond-accuracy aspects. The "Alg." column indicates the classification of algorithms, as described in Sect. 3.3; the "App. domain" column, meanwhile, indicates the application domain of the datasets used for evaluation. The last seven columns indicate whether a study examined certain aspects (Acc. to P-Unbias. in Sect. 3.1), undertook off-line evaluation, and was a user study (Off./U-Std.).

2.4 Terminologies

Target, RecList, and RS: In this paper, the term *Target* indicates the user who is the subject of recommendation. A recommender calculates a list of recommended items for a target. For brevity, we use the word *RecList* and *RS* to indicate a list of recommended items and recommendation system respectively.

Interaction, Love, and Relevance: *Interaction* indicates the *observed* positive actions of users to items (e.g., item purchase logs, positive ratings given by users to items). In this paper, we use the terms *Love* and *Relevance* to indicate the *predicted or hidden* positive attitude of users to items, regardless of the observation of attitude.

3. Survey Results

We describe the details of our survey results, first by explaining the definitions of beyond-accuracy aspects, followed by off-line evaluation metrics. We then classify the recommendation algorithms and describe the idea employed by the algorithms in each classification. Finally, we summarize the user studies carried out by the authors of the surveyed papers. We also explain the relationship between recommendations with regards to beyond-accuracy aspects and other studies in the information-retrieval (IR) domain.

3.1 Definitions of the Aspects

Table 3 summarizes the aspect definitions. A single term is sometimes used in different ways across various papers, to indicate different concepts. For instance, the term "coverage" is used in one group of papers to describe diversity (Table 3). In this paper, we follow the terms shown in Table 3, for consistency of exposition.

Table 3 Aspect definitions

Category	Aspects	Explanation
Primitive	*Accuracy*	The prediction accuracy of the items loved by a target.
	Diversity	The degree of "dissimilarity" among the target items delivered by a single or multiple recommendations. "Measurement" of dissimilarity and "range" of the target items for the dissimilarity calculation vary with the objective or data schema of each RS.
	Novelty	The degree of discovery difficulty of an recommended item by a target. Caused by *non-popularity* or *indifference*.
	Coverage	The amount of distinct items that a RS can recommend. *Prediction coverage* and *catalog coverage* are popular interpretations.
Composite	*Serendipity*	The user satisfaction with the novel items. It can be achieved by a combination of accuracy, novelty and diversity. The importance of each of the aspects to user satisfaction is dependent on subjective perception of each user.
	Popularity unbiasedness	The uniformity of the popularity distribution of the recommended items loved by users.

Accuracy: The term "accuracy" is used to indicate variety of meanings in recommendation studies. For instance, referring to the user satisfaction with generated RecLists [69] or user perception about how well a given RS understands his/her preference [55]. In most recommendation studies, the term accuracy represents *prediction accuracy* (or *objective accuracy*) [29], [55] that indicates recommendation precision for the items preferred by a target user.

Although prediction accuracy is a simpler definition than the others, we need real user participations for evaluation because user preference is a subjective perception. Most of studies could not evaluate their algorithms with real users due to heavy evaluation cost. The studies alternatively adopted log-based *off-line* evaluations. For instance, a given RS is evaluated by the prediction precision for hold-out items already purchased by each user. Due to off-line evaluation, the studies do not evaluate prediction accuracy in the strict sense. Nonetheless in this paper, the term *prediction accuracy* or *accuracy* indicates both of the off-line prediction performance and the strict definition of prediction accuracy to follow the most recommendation studies.

Diversity: Diversity indicates the degree of "dissimilarity" among the target items delivered by a single or multiple recommendations. "Measurement" of dissimilarity and "range" of the target items for the calculation of dissimilarity degree vary with the data schema or objective of each RS.

Three representative types of diversity were found: (1) Simple diversity, *items dissimilar to each other* should be put in the RecList [96]. This interpretation only cares pairwise dissimilarity between each item. (2) Item group representative diversity, *items that represent each of dissimilar groups of similar items in inventory as many as possible* should be placed in the RecList [7]. For instance, a movie recommender focused on (2) recommends movies in a variety of representative genres such as action, comedy, romance in a single RecList while a recommender focused on (1) can recommend movies in various sub-genres in action. (3) User representative diversity, *items that cover the dis-*

similar (diverse) preferences of a given target user should be placed in the RecList [79]. As an example, it is preferable that both sci-fi and romance movies be included in the RecList if the target previously watched movies from the sci-fi and romance genres. For an interpretation from another perspective, Lathia *et al.* [40] introduces temporal diversity, and suggests that recommended items for the same target be changed as time goes by.

Novelty: Novelty refers to the degree of discovery difficulty of an item recommended to a target. Recommenders aware of this aspect should include novel items that might be preferred by the target. Item *nonpopularity* and user *indifference* are two major interpretations of the cause of "novelty" [75]. A user may not know a given item because (1) it is simply not popular enough, or (2) the user does not have an interest in the group of items, irrespective of popularity.

Coverage: Coverage refers the amount of distinct items that a RS can recommend. Roughly two kinds of interpretations about coverage have been discussed in the RS literature. Prediction coverage [29] indicates the ratio of distinct user-item pairs whose *user-item relevance can be predicted (calculated)* by a given RS. Catalog coverage indicates [26], [29] the number of distinct items *actually recommended* for all users. Recommenders that work effectively with regards to this aspect can evaluate user preference to any item in a store inventory or can promote every relevant item in the inventory equally. Items that find it difficult to be known by users obtain a better chance of being recognized. Therefore, coverage is an important aspect not only for users but also for business owners as they aim to make higher profits.

Serendipity: Serendipity indicates that *the user satisfaction with the recommended items that are novel* to the user. [26], [29], [45]. For instance, a documentary film that a user never heard of is a novel item but generally not a serendipitous item if he/she is not interested in documentary films. On the contrary, if the film is matched to the user preference or the user finds out the film is interesting enough to positively change their attitude to the genre to which the user was once indifferent, then the film is a serendipitous item. By definition, this aspect relates to accuracy, novelty, and diversity, but the importance of each aspect to user satisfaction is highly dependent on subjective perception of each user. Owing to this subjectivity, there are only a few studies on this topic.

Popularity unbiasedness: Popularity unbiasedness indicates the *uniformity of the popularity distribution of the recommended items* loved by users [89]. In other words, each item loved by a target should have an equal chance of being recommended, regardless of its popularity. The uniformity differentiates popularity unbiasedness from serendipity. For instance, even if a recommender recommends diverse and novel items that satisfy the target users, it is difficult to say whether the recommender performs well in terms of popularity unbiasedness, if those items occupy only the items with similar popularity.

3.2 Evaluation Metrics

Off-line evaluations that do not require the participation of real users constitute a popular evaluation methodology in the RS literature (Table 2). Recommendations are evaluated by using a set of evaluation metrics, along with RecLists and test data. The test data are extracted from *leave-x-out* user interactions, such as the $x\%$ of item purchase logs that were not used in the algorithm training phase. In this section, we describe the representative evaluation metrics used for each beyond-accuracy aspect. We classify the evaluation metrics according to (1) how many of the aspects described in Sect. 3.1 are measured by a single metric, and (2) the target aspects that each metric tries to measure.

Because many metrics can be grouped into a single classification, we only describe the representative metrics for each classification. We selected the representative metrics, based on the following two criteria. (C1) If there are more than one interpretation about a given aspect described in Sect. 3.1, at least one metric that relates to each interpretation should be selected. (C2) Among the metrics that satisfy (C1), the metric used by the greatest number of surveyed papers is selected.

3.2.1 Single-Aspect Metrics

Single-aspect metrics indicate the metrics that measure performance in a single aspect.

Accuracy: Measures for prediction accuracy have been extensively studied in IR field and adopted by many researchers in recommendation field according to the prediction objectives of RSs. For instance, (1) when the objective is just finding relevant items of a given user, therefore the order of the relevant items is not important, the prediction performance is measured by the ratio of the number of *hit items*—the items found in both the test data of the user and the RecList for the user—to the number of the items in a RecList, or to the number of all items in the user's test data. (e.g., *Precision* or *Recall* [29]). When we want put higher score to the relevant items located in the higher position of a RecList, a measure to put higher scores to higher ranked hit items is used (e.g., *NDCG* [35]). On the other hand, (2) if the objective is predicting the ratings on items given by the target users, the evaluation is done by measuring the error of the predicted ratings on the items in the test data from the actual ratings given by target users on the same items (e.g., *RMSE*, *MAE* [29]). Herlocker *et al.* [29] did extensive survey on accuracy metrics used in RSs. Interested readers also can refer to [59] and [28].

Diversity: An intra-list dissimilarity (*ILD*) measure is widely used to evaluate diversity among recommended items. Equation (1) shows a typical *ILD* form.

$$ILD(R) = \frac{\sum_{r_1 \in R} \sum_{r_2 \in R} dissim(r_1, r_2)}{|R| \cdot |R|}, \tag{1}$$

where R indicates a RecList for a target and $dissim(r_1, r_2)$

is a dissimilarity function that quantifies the difference between two items. Therefore, ILD measures the average pairwise item dissimilarity in a RecList. The function of $dissim(r_1, r_2)$ varies with the objective or underlying data schema of each RS. For instance, Ziegler *et al.* [96] computed the dissimilarity between product sets based upon their taxonomy (e.g, archeology books vs. mathematics books).

S-Recall is another popular metric. *S-Recall* was originally proposed for use in web search result diversification [87]. As shown in Eq. (2), the metric measures the number of retrieved subtopics that relate to a given subject.

$$S - Recall(R) = \frac{|\bigcup_{r \in R} S_r|}{|S|}, \tag{2}$$

where $S_r \subseteq S$ indicates the set of subtopics that relates to item r, and S indicates the set of all subtopics that relates to the given subject. The subtopic set S is defined by the application requirement. If the objective is recommending a diversity of popular genres, then S is equivalent to the set of popular movie genres [7], [77]. If an application aims to cover as much as possible a movie genre watched by a target u, then S is equivalent to the set of the movie genres watched by u [84].

Novelty: As described in Novelty in Sect. 3.1, there are two major interpretations concerning item novelty: the nonpopularity of items, and user indifference to items. Item popularity (IP) [64], [85], [92] measures the average popularity of the items in a recommendation list. A typical IP has a form similar to Eq. (3).

$$IP(R) = \frac{\sum_{r \in R} pop_r}{|R|}, \tag{3}$$

where pop_r indicates the popularity of item r. One way of calculating pop_r is to count the number of user–item interactions observed among the interactions.

Distance-from-user (DU) [88] measures user indifference to the items in a list (Eq. (4)).

$$DU(R, L_u) = \frac{\sum_{r_1 \in R} \sum_{r_2 \in L_u} dissim(r_1, r_2)}{|R| \cdot |L_u|}, \tag{4}$$

where L_u is a set of items with which user u has interacted. The idea behind DU is that if an item is dissimilar from the items with which user u's frequently interacts, then u is highly unlikely to know about the item.

Coverage: Because prediction coverage indicates the ratio of distinct user-item pairs whose user-item relevance can be predicted by a given RS, a simple way of prediction coverage measurement is that selecting random sample of user-item pairs, then measuring the percentage of the pairs whose user preference on item can be calculated by a given RS [29].

Aggregate diversity (Agg-Div) [23] is a simple measure for calculating catalog coverage. As shown in Eq. (5), the metric simply accumulates the number of distinct items in the RecLists of all users. Variations in Agg-Div are found in [26].

$$Agg - Div = \left| \bigcup_{u \in U} R_u \right|, \tag{5}$$

where U indicates the set of all users and R_u indicates the set of the items recommended to user u.

3.2.2 Multi-Aspect Metrics

Multi-aspect metrics measure the combined performance of more than one aspect.

Accuracy–Diversity: *a-NDCG* measures the combined performance of accuracy and diversity. The metric was originally proposed by Clarke *et al.* [19] for measuring the degree of diversification in web search results. *a-NDCG* has a tuning parameter of $a \in [0.0, 1.0]$, which indicates the strength of penalization on the appearance of similar items in a RecList. When the value of a is 0, then *a-NDCG* assigns a full weight to accuracy; therefore, *a-NDCG* is equivalent to *NDCG* [35]. Otherwise, when the value of a is 1, *a-NDCG* assigns a full weight to diversity. In this case, the inclusion of any similar items in the RecList is not rewarded.

Similarly, Vargas *et al.* [75] propose *EILD*, which incorporates the discounted weight calculated from the item position—in addition to the accuracy of the items and intra-list diversity—to assign greater importance to the top-positioned accurate items. Therefore, *EILD* rewards the diversified RecLists that contain more of the interacted items at the tops of lists.

Accuracy–Novelty: Popularity stratified recall (*ps-Recall*) was first proposed by Steck *et al.* [67]. *ps-Recall* is a recall that rewards the retrieval of less-popular relevant items and penalizes the retrieval of popular relevant items.

Vargas *et al.* [75] propose *expected popularity complement (EPC)* and *expected profile distance (EPD)*. Similar to *EILD*, both measurements incorporate the discounted weight calculated from the item position, in addition to the accuracy of the items and item novelty. *EPC* and *EPD* are distinguished by their interpretation of novelty, as described in Novelty in Sect. 3.1: *EPC* measures novelty based on item nonpopularity, and *EPD* measures novelty in terms of user indifference.

Popularity unbiasedness: The *Gini-coefficient* [27], a common measure of distributional inequality, is a widely used evaluation metric. The *Gini-coefficient* indicates the difference between the area under the Lorenz curve and the area under the ideal curve, which indicates the perfect uniform distribution. In the studies of popualrity bias of recommendation, the Lorenz curve $L(x)$ indicates the ratio of recommendation made for the least $100x\%$ ($x \in [0.0, 1.0]$) popular items. Therefore, if recommendations are evenly made for all items, regardless of popularity, the Gini-coefficient value will approach 0; meanwhile, it will approximate 1 if recommendations are concentrated on a few popular items. The *concentration index* [89] is a variation of the Gini-coefficient that considers only the recommendations that could predict

interacted items, and it is also used.

3.3 Recommendation Algorithms

Based on the survey results, we classified the surveyed recommendation algorithms into six groups: multi-objective, modification of CF, clustering-based, graph-based, hybrid, and other algorithms.

3.3.1 Multi-Objective Algorithms

A substantial proportion of the surveyed algorithms are classified as multi-objective algorithms. To recommend items for a user, the algorithms select set R (RecList) of k recommended items that maximize or minimize an objective function, from the candidate item set D of n ($n > k$) items. The objective function is represented by a combination of multiple subobjectives from different aspects. A typical form of the objective function is a weighted linear combination of different subobjectives, as shown in Eq. (6).

$$Obj(u, R) = \sum_{a \in A} w_a \cdot f_a(u, R), \tag{6}$$

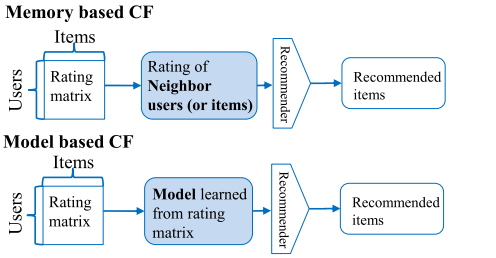
where A indicates a set of aspects; additionally, w_a and $f_a(u, R)$ represent the weight of an aspect a and the subobjective function for a with a target u and a selected item set R, respectively. In most algorithms, one subobjective function indicates the relevance of items in R to u thus related to accuracy, and the other subobjective functions indicate the beyond-accuracy aspects (e.g., diversity or novelty) of the items in R. Generally, the scores of items calculated by conventional recommendation algorithms are used for the subobjective function that relates to accuracy. Intra-list diversity (Eq. (1)) is a popular subobjective for the other aspects. However, other metrics can also be used. For w_a, most of the algorithms require manual tuning; however, some algorithms automatically adjust the weights by using a method such as a genetic program [8], [58].

As [14] points out, deriving the optimal item set R from D is NP-hard, as problems can be reduced to an existing NP-hard problem, such as a set cover or maximum dispersion problem. If all the subobjective functions used are monotonically increasing submodular functions, greedy selection algorithms will guarantee a minimal bound [13]. A typical greedy selection algorithm is shown in Algorithm 1.

Algorithm 1 Greedy selection

procedure GREEDYSELECT
Input: item set D, user u, number $k \leq |D|$
Output: item set R (RecList)
  $R \leftarrow \phi$
  while $|R| < k$ **do**
    $a \leftarrow \arg\min_{r \in D-R} Obj(u, R \cup \{r\})$
    $R \leftarrow R \cup \{a\}$

The candidate item set D is normally generated by taking the n top-ranked items from previously proposed recommendation algorithms that mainly focused on accuracy.



Neighbors (Models) based on the users who have similar user-item interaction pattern.
→ Based on more diverse features from various methodologies.

Fig. 2 Modification in CF methods

Another way of deriving the suboptimal item set R is through the use of binary quadratic programming [31], [34].

3.3.2 Modification of Collaborative Filtering

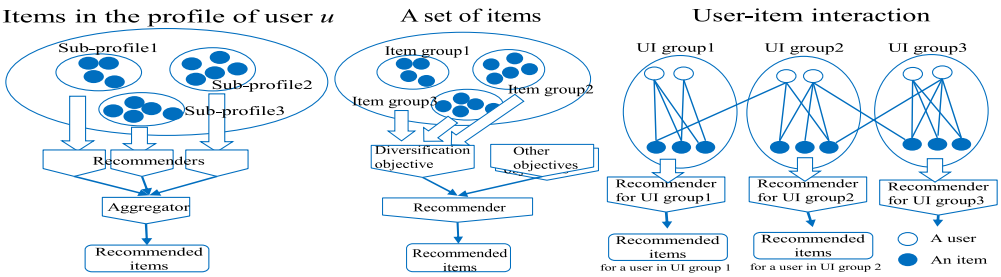
Another stream of research directly modifies CF algorithms to deal with beyond-accuracy aspects.

Memory-based CF (MemCF) and *Model-based CF* (ModelCF) are popular classifications of CF algorithms. As shown in Fig. 2, with MemCF, the full user–item interaction history is stored in memory. MemCF is further classified into User-based CF (UCF) and Item-based CF (ICF). With UCF, users who have interaction patterns similar to those of a target user u are selected as the neighbors of u, and items are recommended based on the neighbors' interactions. With ICF, the term similar "users" as used in UCF are replaced with the term similar "items," and items are recommended based on their similarity to the items with which u interacted. In ModelCF, models that represent each user and item as inferred factors are stored in memory. The interaction patterns of the neighbor users that are similar to those of u are encoded into the inferred factors for each user and item. Then, items are recommended, based on factor similarity between an item and u. Given the mechanism's ease of use, a majority of studies leverage MemCF.

Neighbors/Models with Multi-aspects: Because *similarity* between users is a critical concept in CF, it is natural that many studies would focus on the neighborhood selection or the model inference that incorporates the multiple aspects.

To improve novelty, the discounted weight of popular items in a similarity calculation between users [93] and in model training [67] has been proposed. Shi *et al.* [65] used a dimension reduction technique to determine the representative factors of items in an intrinsically long-tail environment. In an environment where most items have only a few interactions with users, improving recommendation accuracy through dimension reduction is followed by recommending less-popular items more accurately.

For other aspects, Hurley *et al.* [32] proposed a personalized ranking method that incorporates intra-list diversity in model training; Niemann *et al.* [49], meanwhile, adopted a co-occurrence-based similarity function in MemCF to im-



(a) User profile clustering (b) Item clustering (c) Interaction clustering

Fig. 3 Clustering algorithms

prove coverage (aggregate diversity).

Neighbor Filtering: In conventional MemCF, the top n most similar users are selected as neighbors. Some studies create a recommender that can deal with beyond-accuracy aspects by employing less-similar users as neighbors. Adamopoulus *et al.* [2] ordered users in terms of similarity to target u, in descending order; they sampled users from the first user, according to the predefined probability distribution. Chatzicharalampous *et al.* [17] exchanged similar users in the neighbor set who love popular items, instead choosing to use less-similar users who love less-popular items.

Alternative Way of Recommendation: Conventional CFs generally promote the items purchased by multiple similar neighbors. However, there exist studies whose authors took different strategies. Said *et al.* [61] proposed a recommender that recommends the items with which the furthest neighbors infrequently interact. The furthest neighbors indicate the group of users whose purchase pattern is dissimilar to a target. Wu *et al.* [84] tried to cover different aspects of the preference of target u with items in the RecList. This goal was achieved by constructing a neighbor set containing the users who shared many purchased items with u, and then by giving recommendation priority to the items that more of u's neighbors had purchased. Vargas *et al.* [78] improved sales diversity (coverage, popularity unbiasedness) by recommending users to items.

3.3.3 Clustering-Based Algorithms

The idea of clustering algorithms is to divide data into sub-data groups that have different characteristics among them. Generally, the clustering algorithms are used as a preprocess of recommender training. From a clustering subject perspective, three methodologies were gleaned from the surveyed studies.

User Profile Clustering (Fig. 3 (a), [81], [88]): The items in user u's profile are clustered by using a pairwise item similarity. A conventional CF is trained as if each item cluster were equivalent to a single user. When a system calculates a recommendation for user u, subrecommendations are calculated for each cluster of u. Additionally, the subrecommendations are aggregated to generate a recommendation for u. By doing this, the recommendation can incorporate items that represent different subpreferences of u.

Item Clustering (Fig. 3 (b), [10], [42]): The items in the whole item set I are clustered. Because each cluster represents item groups that feature different characteristics, the cluster membership information of the items is used to measure the degree of diversification in diversity-aware recommenders.

Interaction Clustering (Fig. 3 (c), [43]): User–item clusters are calculated by using user–item or item–item similarity, as found in interaction logs. Each cluster represents the users who have similar interaction patterns and the items with which the users frequently interacted. A conventional algorithm is trained for each cluster; this generates recommended items for a target, by using the algorithm trained with the cluster to which the target belongs. By filtering out unrelated popular items, the relative importance of the niche items loved by the members of the group to which the target belongs increases. In this way, we can recommend less-popular but relevant items to targets.

3.3.4 Graph-Based Algorithms

Two groups of graph-based studies are popular in terms of making recommendations while considering beyond-accuracy aspects.

Cost-flow-based Algorithms: Cost-flow-based algorithms [64], [85] constitute an extension of hitting-time-based algorithms. Most of the algorithms are used to recommend novel and relevant items for a given user u.

In a hitting-time-based algorithm, user–item interaction data are visually represented as a bipartite graph, as shown in Fig. 4. When we recommend items for a user or a user group S_u, we recommend k items that have the least amount of hitting time. (Hitting time indicates the expected number of hops needed for a random walk starting from a given item to reach any node in S_u or the items in S_i with which S_u interacts.) Because of the time-reversibility of the first-order Markov chain in Eq. (7), the algorithms prefer less-popular items.

$$\pi_i p_{i,u} = \pi_u p_{u,i} \Leftrightarrow p_{i,u} = \pi_u p_{u,i} / \pi_i \tag{7}$$

π_u represents the initial probability of a random walk, which indicates the probability that a random walk will start from

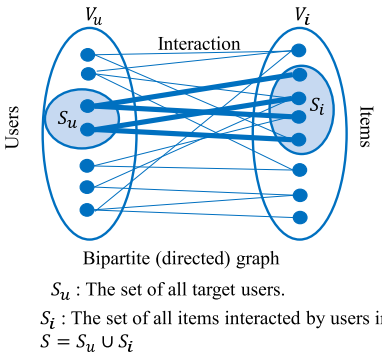


Fig. 4 A bipartite graph representation of rating matrix

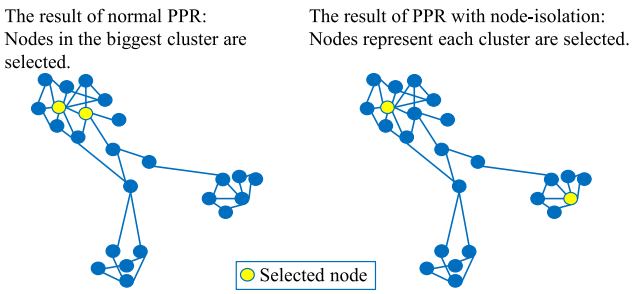


Fig. 5 Results of PPR and PPR with node-isolation

node u; $p_{u,i}$ represents the transition probability of a random walk from user node u to item node i. By dividing each side of Eq. (7) by π_i, we can know that $p_{i,u}$ is inversely proportional to π_i. Because π_i is proportional to the popularity of item i, the inverse proportionality indicates that a less-popular item i has a better chance of reaching the target user u. Therefore, the hitting time of i becomes small if i is a less-popular item.

Hitting time considers the transition probability that is calculated solely from graph topology ($p_{u,i}, p_{i,u}, \pi_u, \pi_i$). The cost flow extends the hitting time by making the transition probability of taking care of the contexts unique to recommendations, such as user–item similarity or item popularity. Specifically, cost flow introduces the concept of transition cost, which is the cost a random walk must pay to transit from node i to node j; this is in addition to transition probability $p_{i,j}$. More valuable transitions in the recommendation context bear lower costs. For example, a transition cost will be low if i and j are relevant, or if j represents a less-popular item, or a user who has a narrow interest. The conditions promote relevant items to users, unpopular items, and users who have an obvious preference for specific items from random walks. *Cost flow* indicates the total cost to reach S_u (or S_i); it is calculated for each item, and k items with the lowest cost flow are recommended. By doing this, more meaningful long-tail items are recommended.

Penalized Popular Users/Items: The algorithms in this category are generally used to calculate more diverse RecLists [39] or more novel RecLists [18].

Küçüktunç *et al.* [39] proposed a node isolation method. A graph $G = (V, E, W)$ is constructed for a target u, and a vertex $v \in V$ indicates an item. An edge $e \in E$ is constructed if a similarity or importance between vertex i and j is above a predefined threshold. Each $e \in E$ has a corresponding weight $w \in W$ that indicates the similarity between two vertices. Personalized page rank (PPR) [24] starts from an arbitrary vertex. However, unlike standard PPR, node isolation-based algorithms detach the vertex that acquires the highest weight from the graph G at some time point; they then proceed with PPR until the next time point at which the next detachment node is selected. This process is continued until k isolated nodes are found, whereupon they are recommended to target u. Because the most connected node and its edges are isolated from the graph, the random walk wandering around the isolated node is significantly reduced. Instead, the random walk goes to the second-most connected node that is not yet isolated. In this way, node isolation-based algorithms can recommend the items that (1) represent each item group and (2) are preferred by the target user u. The right-hand side of Fig. 5 differs from standard PPR results, which tend to recommend items similar to the most popular item (i.e., on the left-hand side of Fig. 5).

Christoffel *et al.* [18] proposed a random walk method that penalizes popular nodes. In their method, a graph $G = (V, E, W)$ represents observed user–item interactions. A vertex $v \in V$ indicates a user or item. Edge $e \in E$ is constructed if there is an interaction between user vertex u and item vertex i. When the method calculates a node's weight based on the visit frequency of the random walk, the weight of a popular node is discounted inverse-proportionally to the degree (i.e., popularity) of the node. Therefore, less-popular nodes can acquire more weight according to their method, relative to standard random-walk-based weighting algorithms.

3.3.5 Hybrid Algorithms

Hybrid algorithms combine the results from existing recommendation algorithms that bear different characteristics or perspectives, in order to calculate RecLists [12], [30], [46], [83], [92], [94]. Among the surveyed studies, we found a number of combined algorithms; of particular interest was a novelty prediction model inferred from a novelty matrix, where the perceived item novelty is explicitly given by users as one of its elements [30]. Another was a user preference model on noncontent preference attributes such as popularity and item recency [46]. All of the surveyed studies used a combination of scores or rankings given for item i, as derived by different algorithms, to make a final recommendation to a given user.

3.3.6 Other Algorithms

There exist algorithms that do not "fit" into the aforementioned typology. Ishikawa *et al.* [33] exploited information diffusion theory to recommend novel technical articles to interested users; additionally, Shi *et al.* [66] used portfolio theory from the field of economics to tune recommendations

Table 4 A summary of user studies

Category	Method	Employment type	Paper (# of users)
On-line experiment	Measuring the indexes of interest	N/A	[70](5000~40000)
			[72](100000)
Small sized user study	Questionnaire	On-line volunteers	[61](132), [41](26), [96](2125), [21](582)
		Cloud sourcing	[6](200, 459), [1](132)
		Off-line	[92](21), [85](50), [42](50), [30](40)
		Unknown	[71](25), [67](20), [10](50), [16](288), [20](382)

between risk-taking and risk-aversion.

Kapoor *et al.* [36] tried to predict user perceptual demand for novelty at the current time, by using a logistic regression to propose a time-based, novelty-aware recommendation; Szpektor *et al.* [70] achieved a diversified recommendation set by repeatedly sampling subpreferences from a user preference tree and recommending items relevant to the sampled subpreferences.

Kohli *et al.* [37] studied on-line result diversification algorithms by using a multi-arm bandit algorithm. Zhang and Hurley [90] propose the use of a statistical model in making diversified recommendations.

3.4 User Studies

User studies indicate evaluations that feature real user participation. Although user studies incur higher costs than off-line evaluations, the results indicate that user preferences more closely approximate true user perceptions. For instance, Rosseti *et al.* [60] and Garcin *et al.* [25] reported the accuracy of methods evaluated in off-line evaluations was reversed, compared to those of user studies. Because the performance of most beyond-accuracy aspects are highly dependent on the subjectivity of user perception, user studies have grown in importance. 17 of the 72 surveyed papers carried out user studies. Table 4 provides a summary.

3.4.1 On-Line Experiments

On-line experiments refer to evaluations carried out on real service provision. The currently running algorithm is replaced by a comparison algorithm, and then the indexes of interest (e.g., click-through rate) are collected and compared to those of the replaced algorithm. However, only two of the 17 user studies had carried out on-line experiments; this is because an on-line experiment is a high-risk experiment, given the opacity of the performance of the proposed algorithms. In addition, most researchers in academia do not provide on-line services.

3.4.2 Small-Sized User Studies

Participant Employment: Most of the user studies carried out evaluations that featured a small user sample. Given the use of such small user samples, questionnaire-based studies have been popular. Participant employment tends to be roughly classified into three types: (1) *Off-line* employment, (2) *On-line volunteers*, and (3) *Cloud sourcing*. Off-line employment is a traditional method. In off-line employment, participants visit a designated place and participate the experiment by meeting face-to-face with an experiment conductor. On-line volunteers are participants who voluntarily responded to a call for participants, via e-mail or post mail. The number of participants often depends on the credibility of the experiment conductor: if the conductor can obtain the cooperation of a real service provider, a relatively large number of users tend to become involved [21], [96]. Recently, a decent number of users can be gathered through cloud sourcing, such as that through Amazon Mechanical Turk (https://www.mturk.com). Unlike with off-line employment, on-line volunteers and cloud-sourced workers perform evaluation tasks in remote places, far from where the conductor is. Therefore, for these types of studies, it is better to prepare a mechanism that guarantees evaluation quality [61].

Questionnaire: All of the 15 studies to feature small user samples had used a questionnaire as an evaluation method—likely on account of sample size. A typical questionnaire-based evaluation features the following process. (1) The recommenders that are the topics of the evaluation are trained for each user by using information provided by the user. For instance, a user is asked to rate at least 10 movies from the top 100 most popular movies, from among all the genres in the system. Because the purpose of training is to learn about exact user preferences, sufficiently popular items are provided as selection candidates. (2) Then, each recommender provides the users with a top-k recommendation list. The lists are randomly showed to users, to eliminate some bias that arises from the order or position of display. (3) After users examine the lists, a questionnaire is shown and the users reply to the questionnaire.

A questionnaire consists of questions used to measure performance in terms of the aspects of interest, and the questions ask about overall user satisfaction. For example, if an interest is to improve serendipity by recommending novel items, then the questions that ask about the novelty of the recommended item and those that ask about satisfaction with the novel items will appear alongside those questions that ask about the overall satisfaction of the recommender. Depending on the nature of each question, users can provide answers on a five-point scale, or with binary yes/no ratings. The results are summarized and compared between the recommenders. ANOVA and t-tests are used to test the statistical significance of the results with respect to the evaluation design.

3.4.3 Results

Many user study results indicated that overall satisfaction with the beyond-accuracy aspect adopted recommenders is higher than that with the accuracy-focused recommenders [67], [70], [84], even in cases where the accuracy of the beyond-accuracy adopted recommenders is lower than

that of competitors [10], [30], [61], [85], [96]. However, it is important to strike an appropriate balance between accuracy and beyond-accuracy. For example, in the studies of [96] and [67], an appropriate balance between accuracy and diversity helped achieve the best user satisfaction results. One possible reason is that accurate items recommended by conventional algorithms are kinds of popular items that are recognizable by users. Recognition plays an important role in user trust with regards to recommendations, by letting users judge their preferences for the recommended items. In other words, users cannot assess a RecList that contains fully diverse and novel items, as they will not know the items well, and the items will not appear to follow any reasonable selection rule.

Some of the surveyed studies were user studies, as there was no alternative way for their authors to evaluate concepts through the use of off-line evaluations [41], [71]. Other studies that did not carry out user studies showed better accuracy by using off-line evaluations with their dataset [81], or they supposed a situation in which beyond-accuracy aspects become critical—whereupon they propose a way of introducing beyond-accuracy aspects for the supposed situation [3], [49]. However, in both cases, user satisfaction with the proposed recommendation was not clear, and user satisfaction supposedly varied with respect to the application purpose. Therefore, user studies represent an important evaluation methodology, especially in the context of recommendations that adopt beyond-accuracy aspects.

3.5 Relation to Other Information-Retrieval Domains

Beyond-accuracy aspects have also been studied in other information-retrieval fields. Especially, many discussions and ideas with regards to web search result diversification have been influenced on recommendation algorithms. In web search result diversification, extrinsic diversity indicates the interpretation diversity of queries; intrinsic diversity indicates the aspect diversity for the same interpretation [57].

If a user issues an ambiguous query that can garner many different interpretations, it is difficult for search engines to generate results that match the exact intent of the user. For example, the query "Jaguar" could refer either to an animal species or a car brand; in such a case, it is better to include in the search results at least one document that relates to each interpretation, to mitigate the risk that some users will not obtain any documents that relate to their intent. Even if a search engine knows that the user wants information about cars, adding a document that contains duplicated information is not valuable (e.g., adding a specification document for the car, when similar documents are already included in the list). The two aforementioned forms of diversity are similar to diversity and novelty, in terms of making recommendations.

What sets the beyond-accuracy aspect recommendations apart from searched document diversification is the nonexistence of explicit queries. Instead of queries, the subject of diversification is an individual user. Compared to search logs—which can be used to infer query interpretations—a user does not have sufficient interaction logs that can be used to infer diverse user preferences. Therefore, the so-called cold-start problem needs to be addressed. In addition, the degree of diversification is user-dependent. Searched document diversification focuses on a utility that can maximize the number of users who can find at least one relevant document; therefore, it mainly studies global diversity. In contrast, a recommender should predict as much as possible a diversity of items loved by a target user; recommendations should therefore put more weight on personal diversity.

Another difference is the role of novelty. In searched document diversification, novelty relates to intrinsic diversity; this means that a document contains novel (different) information about the same interpretation, from documents already inserted in the results list. If documents contain similar novel information, then selecting the most popular document is better, because a popular document normally contains high-quality information. However, in making recommendations, the matter becomes more complicated: because popular items are more likely to be known to users, acquiring an appropriate novelty degree depends on the individual users.

IA-Select [4] and xQuAD [62] are document diversification methods based on MMR-style multi-objective algorithms [13]. The thinking behind the algorithms has been introduced to the recommendation field, through several studies [7], [76], [79].

A hitting-time-based algorithm [44] proposed for query suggestion diversification was extended to cost-flow-based algorithms [64], [85]. Additionally, a node isolation-based algorithm [95] proposed for document summarization was imported for use in making recommendations [38].

Wang *et al.* [80] introduced portfolio theory, as found in the field of economics, to the field of information retrieval. Portfolio theory provides a framework by which to tune the degree of risk inherent in the retrieved information. The portfolio concept introduced by Wang *et al.* was extended to the study of recommendations by Shi *et al.* [66].

4. Future Directions

Although many recommender algorithms have been proposed for beyond-accuracy aspects, there remain many open questions and challenges to be studied. In this section, we summarize future research directions that were identified in the course of this survey.

Implementation of Beyond-accuracy Aspects: Despite there being simple abstract definitions of beyond-accuracy aspects, actual interpretations and implementations vary by application. For example, the genres of movies watched by a target are diversified among movie recommendations, and representative articles describing a given topic are diversified among news recommendations. Even if the interpretation of aspects were similar, the aspect measurement func-

tion and the algorithm can also differ with regards to data availability and algorithm type. Although many algorithms have been proposed, there is some consensus regarding the frameworks that can be applied to various applications.

Personalization: As we reported in Sect. 3.3, many studies used a weighted combination of scores that represent various aspects. However, the combination weights in most studies are global in nature: the same weights are applied to calculate a RecList for every user. Although some studies make use of personalized combination weights [6], [72], it is difficult to infer appropriate weights, as the dataset available from each individual user is small. In addition, it is not clear how best to address the cold-start problem. Even in research on personalized combinations, data that exceed some quantity threshold are needed to infer personalized combination weights [6]. How best to select appropriate weights for users for whom there are insufficient data remains unclear. Another problem is that, in most studies, personalization is covered by the accuracy aspect. With the other aspects—such as diversity and novelty—only a few methods [50], [72], [79] pay attention to personalized diversity or novelty.

Correlation between Off-line Performance and User Perception: Given the highly subjective nature of beyond-accuracy aspects, many of the surveyed studies listed, side by side, the results of off-line performance (measured in terms of the metrics described in Sect. 3.2). Knowledge of correlation, if any, between off-line performance and real user perception would be useful in making inferences about user satisfaction from off-line performance information. However, to the best of our knowledge, no study has examined this correlation.

Explanation: As discussed in Sect. 3.4.3, explicable recommended items are important to recommenders aware of the beyond-accuracy aspects, because explanations as to why the items are recommended can provide users with information by which to assess whether unfamiliar items in their RecLists fulfill their preferences. However, explicable beyond accuracy adopted recommendations have not broadly studied, in spite of their importance.

Interface and Layout: Generally, accuracy and beyond-accuracy aspects are in a trade-off. Therefore, in some cases, it would be a good idea to calculate different RecLists from recommenders that feature different aspects, and display those RecLists concurrently. In such cases, an appropriate layout and interface by which to navigate the items within the RecLists would be needed, to achieve user convenience and satisfaction. Among the studies we surveyed, a few (e.g., [72]) touch upon this issue. Clearly, interface and layout design remains another domain to be investigated by future research.

5. Conclusion

In this study, we surveyed recommendation systems that address beyond-accuracy aspects. We described an algorithm classification typology, various evaluation metrics, and some commonly employed user study methodologies. Our results indicate that while the recommendations garnered by addressing beyond-accuracy aspects have improved to some extent, there remains considerable room for improvement.

References

[1] S. Abbar, S. Amer-Yahia, P. Indyk, and S. Mahabadi, "Real-time recommendation of diverse related articles," Proc. WWW '13, pp.1–12, 2013.
[2] P. Adamopoulos and A. Tuzhilin, "On over-specialization and concentration bias of recommendations: probabilistic neighborhood selection in collaborative filtering systems," Proc. RecSys'14, pp.153–160, 2014.
[3] G. Adomavicius and Y. Kwon, "Improving aggregate recommendation diversity using ranking-based techniques," IEEE Trans. Knowl. Data Eng., vol.24, no.5, pp.896–911, 2012.
[4] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong, "Diversifying search results," Proc. the 2nd ACM Int'l Conf. on Web Search and Data Mining, pp.5–14, 2009.
[5] C. Anderson, "The long tail: why the future of business is selling less of more," Proc. Hyperion, New York, 2006.
[6] A. Ashkan, B. Kveton, S. Berkovsky, and Z. Wen, "Optimal greedy diversity for recommendation," Proc. the 24th Int'l Conf. on AI, pp.1742–1748, 2015.
[7] F.M. Belém, R. Santos, J. Almeida, and M.A. Gonçalves, "Topic diversity in tag recommendation," Proc. RecSys'13, pp.141–148, 2013.
[8] F.M. Belém, E.F. Martins, J.M. Almeida, and M.A. Gonçalves, "Exploiting novelty and diversity in tag recommendation," Proc. European Conf. on Info. Retrieval 2013, pp.380–391, 2013.
[9] I. Benouaret and D. Lenne, "A package recommendation framework for trip planning activities," Proc. RecSys'16, pp.203–206, 2016.
[10] R. Boim, T. Milo, and S. Novgorodov, "Diversification and refinement in collaborative filtering recommender," Proc. the 20th ACM Int'l Conf. on Info. and knowledge management, pp.739–744, 2011.
[11] K. Bradley and B. Smyth, "Improving recommendation diversity," Proc. 12th Irish Conf. on AI and Cognitive Science, pp.85–94, 2001.
[12] D.G. Bridge and A. Ferguson, "Diverse product recommendations using an expressive language for case retrieval," Proc. the 6th European Conf. on Advances in Case-Based Reasoning, pp.43–57, 2002.
[13] J. Carbonell and J. Goldstein, "The use of MMR, diversity-based reranking for reordering documents and producing summaries," Proc. SIGIR'98, pp.335–336, 1998.
[14] B. Carterette, "An analysis of NP-completeness in novelty and diversity ranking," Proc. Info. Retr., vol.14, no.1, pp.89–106, 2011.
[15] P. Castells and S. Vargas, "Novelty and diversity metrics for recommender systems: choice, discovery and relevance," Proc. Int'l Workshop on Diversity in Document Retrieval, pp.29–37, 2011.
[16] Ó. Celma and P. Herrera, "A new approach to evaluating novel recommendations," Proc. RecSys'08, pp.179–186, 2008.
[17] E. Chatzicharalampous, Z. Christos, and A. Vakali, "Exploriometer: leveraging personality traits for coverage and diversity aware recommendations," Proc. WWW'15 Companion, pp.1463–1468, 2015.
[18] F. Christoffel, B. Paudel, C. Newell, and A. Bernstein, "Blockbusters and wallflowers: accurate, diverse, and scalable recommendations with random walks," Proc. RecSys'15, pp.163–170, 2015.
[19] C.L.A. Clarke, M. Kolla, G.V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon, "Novelty and diversity in information retrieval evaluation," Proc. SIGIR'08, pp.659–666, 2008.
[20] P. Cremonesi, F. Garzotto, R. Pagano, and M. Quadrana, "Recommending without short head," Proc. WWW'14 Companion, pp.245–246, 2014.
[21] M.D. Ekstrand, F.M. Harper, M.C. Willemsen, and J.A. Konstan,

"User perception of differences in recommender algorithms," Proc. RecSys'14, pp.161–168, 2014.
[22] D.M. Fleder and K. Hosanagar, "Recommender systems and their impact on sales diversity," Proc. the 8th ACM Conf. on Electronic commerce, pp.192–199, 2007.
[23] D. Fleder and K. Hosanagar, "Blockbuster culture's next rise or fall: the impact of recommender systems on sales diversity," Proc. Management Science, vol.55, no.5, pp.697–712, 2009.
[24] D. Fogaras, B. Rácz, K. Csalogány, and T. Sarlós, "Towards scaling fully personalized PageRank: algorithms, lower bounds, and experiments," Proc. Internet Mathematics, vol.2, no.3, pp.333–358, 2005.
[25] F. Garcin, B. Faltings, O. Donatsch, A. Alazzawi, C. Bruttin, and A. Huber, "Offline and online evaluation of news recommender systems at swissinfo.ch," Proc. of RecSys'14, pp.169–176, 2014.
[26] M. Ge, C. Delgado-Battenfeld, and D. Jannach, "Beyond accuracy: evaluating recommender systems by coverage and serendipity," Proc. RecSys'10, pp.257–260, 2010.
[27] C. Gini, "Concentration and dependency ratios (1909 in Italian)," Proc. English translation in Rivista di Politica Economica, vol.87, pp.769–789, 1997.
[28] A. Gunawardana and G. Shani, "A survey of accuracy evaluation metrics of recommendation tasks," The Journal of Machine Learning Research, vol.10, pp.2935–2962, 2009.
[29] J.L. Herlocker, J.A. Konstan, L.G. Terveen, and J.T. Riedl, "Evaluating collaborative filtering recommender systems," Proc. ACM Trans. Info. Systems, vol.22, no.1, pp.5–53, 2004.
[30] Y. Hijikata, T. Shimizu, and S. Nishida, "Discovery-oriented collaborative filtering for improving user satisfaction," Proc. the 14th Int'l Conf. on Intell. user interfaces, pp.67–76, 2009.
[31] N. Hurley and M. Zhang, "Novelty and diversity in top-n recommendation – analysis and evaluation," ACM Trans. Internet Technol., vol.10, no.4, pp.14:11–14:30, 2011.
[32] N.J. Hurley, "Personalised ranking with diversity," Proc. RecSys'13, pp.379–382, 2013.
[33] M. Ishikawa, P. Geczy, N. Izumi, and T. Yamaguchi, "Long tail recommender utilizing information diffusion theory," Proc. the 2008 IEEE/WIC/ACM Int'l Conf. on Web Intell. and Intell. Agent Technol., vol.01, pp.785–788, 2008.
[34] T. Jambor and J. Wang, "Optimizing multiple objectives in collaborative filtering," Proc. RecSys'10, pp.55–62, 2010.
[35] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of ir techniques," ACM Transactions on Information Systems, vol.20, no.4, pp.422–446, 2002.
[36] K. Kapoor, V. Kumar, L. Terveen, J.A. Konstan, and P. Schrater, "I like to explore sometimes: adapting to dynamic user novelty preferences," Proc. RecSys'15, pp.19–26, 2015.
[37] P. Kohli, M. Salek, and G. Stoddard, "A fast bandit algorithm for recommendations to users with heterogeneous tastes," Proc. the 27th AAAI Conf. on AI., pp.1135–1141, AAAI Press, 2013.
[38] O. Küçüktunç, E. Saule, K. Kaya, and Ü.V. Çatalyürek, "Diversified recommendation on graphs: pitfalls, measures, and algorithms," Proc. WWW'13, pp.715–726, 2013.
[39] O. Küçüktunç, E. Saule, K. Kaya, and Ü.V. Çatalyürek, "Diversifying citation recommendations," ACM Trans. Intell. Syst. Technol., vol.5, no.4, article 55, 2015.
[40] N. Lathia, S. Hailes, L. Capra, and X. Amatriain, "Temporal diversity in recommender systems," Proc. SIGIR '10, pp.210–217, 2010.
[41] K. Lee and K. Lee, "My head is your tail: applying link analysis on long-tailed music listening behavior for music recommendation," Proc. RecSys'11, pp.213–220, 2011.
[42] L. Li, D. Wang, T. Li, D. Knox, and B. Padmanabhan, "SCENE: a scalable two-stage personalized news recommendation system," Proc. SIGIR '11, pp.125–134, 2011.
[43] X. Li and T. Murata, "Multidimensional clustering based collaborative filtering approach for diversified recommendation," Proc. 7th Int'l Conf. on Computer Science & Education, pp.905–910, 2012.
[44] H. Ma, M.R. Lyu, and I. King, "Diversifying query suggestion results," Proc. the 24th AAAI Conf. on AI, pp.1399–1404, 2010.
[45] S.M. McNee, J. Riedl, and J.A. Konstan, "Being accurate is not enough: how accuracy metrics have hurt recommender systems," Proc. CHI '06 Extended Abstracts on Human Factors in Computing Systems, pp.1097–1101, 2006.
[46] F. Mourão, L. Rocha, J.A. Konstan, and W. Meira, Jr., "Exploiting non-content preference attributes through hybrid recommendation method," Proc. RecSys'13, pp.177–184, 2013.
[47] T. Murakami, K. Mori, and R. Orihara, "Metrics for evaluating the serendipity of recommendation lists," Proc. the 2007 Conf. on New frontiers in AI, pp.40–46, 2007.
[48] T.T. Nguyen, P.-M. Hui, F.M. Harper, L. Terveen, and J.A. Konstan, "Exploring the filter bubble: the effect of using recommender systems on content diversity," Proc. WWW'14, pp.677–686, 2014.
[49] K. Niemann and M. Wolpers, "A new collaborative filtering approach for increasing the aggregate diversity of recommender systems," Proc. KDD'13, pp.955–963, 2013.
[50] T.D. Noia, V.C. Ostuni, J. Rosati, P. Tomeo, and E.D. Sciascio, "An analysis of users' propensity toward diversity in recommendations," Proc. RecSys'14, pp.285–288, 2014.
[51] J. Oh, S. Park, H. Yu, M. Song, and S.-T. Park, "Novel recommendation based on personal popularity tendency," Proc. the 2011 IEEE 11th Int'l Conf. on Data Mining, pp.507–516, 2011.
[52] U. Panniello, A. Tuzhilin, and M. Gorgoglione, "Comparing context-aware recommender systems in terms of accuracy and diversity," User Modeling and User-Adapted Interaction, vol.24, no.1, pp.35–65, 2014.
[53] S.A.P. Parambath, N. Usunier, and Y. Grandvalet, "A coverage-based approach to recommendation diversity on similarity graph," Proc. RecSys'16, pp.15–22, 2016.
[54] Y. Park and A. Tuzhilin, "The long tail of recommender systems and how to leverage it," Proc. RecSys'08, pp.11–18, 2008.
[55] P. Pu, L. Chen, and R. Hu, "Evaluating recommender systems from the user's perspective: survey of the state of the art," User Modeling and User-Adapted Interaction, vol.22, no.4-5, pp.317–355, 2012.
[56] L. Qin and X. Zhu, "Promoting diversity in recommendation by entropy regularizer," Proc. the 23rd Int'l joint Conf. on AI, pp.2698–2704, 2013.
[57] F. Radlinski, P.N. Bennett, B. Carterette, and T. Joachims, "Redundancy, diversity and interdependent document relevance," Proc. SIGIR Forum 43, pp.46–52, 2009.
[58] M.T. Ribeiro, A. Lacerda, A. Veloso, and N. Ziviani, "Pareto-efficient hybridization for multi-objective recommender systems," Proc. RecSys'12, pp.19–26, 2012.
[59] F. Ricci, L. Rokach, B. Shapira, and P.B. Kantor, Recommender systems handbook, Springer-Verlag New York, Inc., USA, 2010.
[60] M. Rossetti, F. Stella, and M. Zanker, "Contrasting offline and online results when evaluating recommendation algorithms," Proc. RecSys'16, pp.31–34, 2016.
[61] A. Said, B. Fields, B.J. Jain, and S. Albayrak, "User-centric evaluation of a k-furthest neighbor collaborative filtering recommender algorithm," Proc. the 2013 Conf. on Computer supported cooperative work, pp.1399–1408, 2013.
[62] R.L.T. Santos, C. Macdonald, and I. Ounis, "Exploiting query reformulations for web search result diversification," Proc. WWW'10, pp.881–890, 2010.
[63] M. Servajean, E. Pacitti, S. Amer-Yahia, and P. Neveu, "Profile diversity in search and recommendation," Proc. WWW'13 Companion, pp.973–980, 2013.
[64] L. Shi, "Trading-off among accuracy, similarity, diversity, and long-tail: a graph-based recommendation approach," Proc. RecSys'13, pp.57–64, 2013.
[65] K. Shi and K. Ali, "GetJar mobile application recommendations with very sparse datasets," Proc. KDD'12, pp.204–212, 2012.
[66] Y. Shi, X. Zhao, J. Wang, M. Larson, and A. Hanjalic, "Adaptive diversification of recommendation results via latent factor portfolio," Proc. SIGIR'12, pp.175–184, 2012.

[67] H. Steck, "Item popularity and recommendation accuracy," Proc. RecSys'11, pp.125–132, 2011.
[68] R. Su, L. Yin, K. Chen, and Y. Yu, "Set-oriented personalized ranking for diversified top-n recommendation," Proc. RecSys'13, pp.415–418, 2013.
[69] K. Swearingen and R. Sinha, "Beyond Algorithms: An HCI Perspective on Recommender Systems," Proc. SIGIR workshop 2001, pp.393–408, 2001.
[70] I. Szpektor, Y. Maarek, and D. Pelleg, "When relevance is not enough: promoting diversity and freshness in personalized question recommendation," Proc. WWW'13, pp.1249–1260, 2013.
[71] M. Taramigkou, E. Bothos, K. Christidis, D. Apostolou, and G. Mentzas, "Escape the bubble: guided exploration of music preferences for serendipity and novelty," Proc. RecSys'13, pp.335–338, 2013.
[72] C.H. Teo, H. Nassif, D. Hill, S. Srinivasan, M. Goodman, V. Mohan, and S.V.N. Vishwanathan, "Adaptive, personalized diversity for visual discovery," Proc. RecSys'16, pp.35–38, 2016.
[73] I. Fernández-Tobías, P. Tomeo, I. Cantador, T.D. Noia, and E.D. Sciascio, "Accuracy and diversity in cross-domain recommendations for cold-start users with positive-only feedback," Proc. RecSys'16, pp.119–122, 2016.
[74] H. Tong, J. He, Z. Wen, R.i Konuru, and C.-Y. Lin, "Diversified ranking on large graphs: an optimization viewpoint," Proc. KDD'11, pp.1028–1036, 2011.
[75] S. Vargas and P. Castells, "Rank and relevance in novelty and diversity metrics for recommender systems," Proc. RecSys'11, pp.109–116, 2011.
[76] S. Vargas, P. Castells, and D. Vallet, "Intent-oriented diversity in recommender systems," Proc. SIGIR'11, pp.1211–1212, 2011.
[77] S. Vargas and P. Castells, "Exploiting the diversity of user preferences for recommendation," Proc. the 10th Conf. on Open Research Areas in Info. Retrieval, pp.129–136, 2013.
[78] S. Vargas and P. Castells, "Improving sales diversity by recommending users to items," Proc. RecSys'14, pp.145–152, 2014.
[79] S. Vargas, L. Baltrunas, A. Karatzoglou, and P. Castells, "Coverage, redundancy and size-awareness in genre diversity for recommender systems," Proc. RecSys'14, pp.209–216, 2014.
[80] J. Wang and J. Zhu, "Portfolio theory of information retrieval," Proc. SIGIR'09, pp.115–122, 2009.
[81] C. Wartena and M. Wibbels, "Improving tag-based recommendation by topic diversification," Proc. the 33rd European Conf. on Advances in info. retrieval, pp.43–54, 2011.
[82] J. Wasilewski and N. Hurley, "Intent-aware diversification using a constrained PLSA," Proc. RecSys'16, pp.39–42, 2016.
[83] L.-T. Weng, Y. Xu, Y. Li, and R. Nayak, "Improving recommendation novelty based on topic taxonomy," Proc. Workshops on Web Intell. and Intell. Agent Technol., on IEEE/WIC/ACM Int'l Conf., pp.115–118, 2007.
[84] L. Wu, Q. Liu, E. Chen, N.J. Yuan, G. Guo, and X. Xie, "Relevance meets coverage: a unified framework to generate diversified recommendations," ACM Trans. Intell. Syst. Technol., vol.7, no.3, Article 39, 2016.
[85] H. Yin, B. Cui, J. Li, J. Yao, and C. Chen, "Challenging the long tail recommendation," Proc. VLDB Endow., vol.5, no.9, pp.896–907, 2012.
[86] C. Yu, L. Lakshmanan, and S. Amer-Yahia, "It takes variety to make a world: diversification in recommender systems," Proc. the 12th Int'l Conf. on Extending DB Technol.: Advances in DB Technol., pp.368–378, 2009.
[87] C.X. Zhai, W.W. Cohen, and J. Lafferty, "Beyond independent relevance: methods and evaluation metrics for subtopic retrieval," Proc. SIGIR'03, pp.10–17, 2003.
[88] M. Zhang and N. Hurley, "Novel item recommendation by user profile partitioning," Proc. of the 2009 IEEE/WIC/ACM Int'l Conf. on Web Intell. and Intell. Agent Technol., vol.01, pp.508–515, 2009.
[89] M. Zhang and N. Hurley, "Evaluating the diversity of top-n recommendations," Proc. the 21st Int'l Conf. on Tools with AI, pp.457–460, 2009.
[90] M. Zhang and N. Hurley, "Statistical modeling of diversity in top-n recommender systems," Proc. of the 2009 IEEE/WIC/ACM Int'l Joint Conf. on Web Intell. and Intell. Agent Technol., vol.1, pp.490–497, 2009.
[91] M. Zhang and N. Hurley, "Niche product retrieval in top-n recommendation," Proc. the 2010 IEEE/WIC/ACM Int'l Conf. on Web Intell. and Intell. Agent Technol., pp.74–81, 2010.
[92] Y.C. Zhang, D.Ó. Séaghdha, D. Quercia, and T. Jambor, "Auralist: introducing serendipity into music recommendation," Proc. the 5th ACM Int'l Conf. on Web search and data mining, pp.13–22, 2012.
[93] X. Zhao, Z. Niu, and W. Chen, "Opinion-based collaborative filtering to solve popularity bias in recommender systems," Proc. Int'l Conf. on Database and Expert Systems Applications, pp.426–433, 2013.
[94] T. Zhou, Z. Kuscsik, J.-G. Liu, M. Medo, J.R. Wakeling, and Y.-C. Zhang, "Solving the apparent diversity-accuracy dilemma of recommender systems," Proc. PNAS, vol.107, no.10, pp.4511–4515, 2010.
[95] X. Zhu, A.B. Goldberg, J.V. Gael, and D. Andrezejewski, "Improving diversity in ranking using absorbing random walks," Proc. HLT-NAACL, pp.97–104, 2007.
[96] C.N. Ziegler, S.M. McNee, J.A. Konstan, and G. Lausen, "Improving recommendation lists through topic diversification," Proc. the 14th Int'l Conf. on World Wide Web, pp.22–32, 2005.

Jungkyu Han is a Ph.D. student at Waseda Univerisity. His research interests include data mining, artificial intelligence, distributed computing.

Hayato Yamana has been a Professor of Faculty of Science and Engineering at Waseda University since 2005. He is a member of IEEE and ACM and a senior member of IEICE.