Diversity, Serendipity, Novelty, and Coverage: A Survey and Empirical Analysis of Beyond-Accuracy Objectives in Recommender Systems

MARIUS KAMINSKAS and DEREK BRIDGE, Insight Centre for Data Analytics,
University College Cork, Ireland

What makes a good recommendation or good list of recommendations?

Research into recommender systems has traditionally focused on accuracy, in particular how closely the recommender's predicted ratings are to the users' true ratings. However, it has been recognized that other recommendation qualities—such as whether the list of recommendations is diverse and whether it contains novel items—may have a significant impact on the overall quality of a recommender system. Consequently, in recent years, the focus of recommender systems research has shifted to include a wider range of "beyond accuracy" objectives.

In this article, we present a survey of the most discussed beyond-accuracy objectives in recommender systems research: diversity, serendipity, novelty, and coverage. We review the definitions of these objectives and corresponding metrics found in the literature. We also review works that propose optimization strategies for these beyond-accuracy objectives. Since the majority of works focus on one specific objective, we find that it is not clear how the different objectives relate to each other.

Hence, we conduct a set of offline experiments aimed at comparing the performance of different optimization approaches with a view to seeing how they affect objectives other than the ones they are optimizing. We use a set of state-of-the-art recommendation algorithms optimized for recall along with a number of reranking strategies for optimizing the diversity, novelty, and serendipity of the generated recommendations. For each reranking strategy, we measure the effects on the other beyond-accuracy objectives and demonstrate important insights into the correlations between the discussed objectives. For instance, we find that rating-based diversity is positively correlated with novelty, and we demonstrate the positive influence of novelty on recommendation coverage.

CCS Concepts: • **Information systems** → **Information retrieval diversity**; **Novelty in information retrieval**; **Recommender systems**; *Retrieval effectiveness;*

Additional Key Words and Phrases: Evaluation metrics, beyond accuracy, diversity, serendipity, novelty, coverage

ACM Reference Format:
Marius Kaminskas and Derek Bridge. 2016. Diversity, serendipity, novelty, and coverage: A survey and empirical analysis of beyond-accuracy objectives in recommender systems. ACM Trans. Interact. Intell. Syst. 7, 1, Article 2 (December 2016), 42 pages.
DOI: http://dx.doi.org/10.1145/2926720

1. INTRODUCTION

Traditionally, the focus of recommender systems (RS) research has been the accurate prediction of users' ratings for unseen items. However, accuracy is not the only

The reviewing of this article was managed by associate editor Bamshad Mobasher. This work is supported by the Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289.
Authors' addresses: M. Kaminskas and D. Bridge, Insight Centre for Data Analytics, Western Gateway Building, University College Cork, Western Road, Cork, Ireland; emails: {marius.kaminskas, derek.bridge}@insight-centre.org.

DOI: http://dx.doi.org/10.1145/2926720

important objective of recommendation [McNee et al. 2006]. In recent years, the focus of RS research has shifted to such objectives as correctly ranking a set of items (known as the learning-to-rank problem [Shi et al. 2010]) as well as ensuring that the set of recommended items is diverse [Vargas et al. 2014] and that it contains novel items [Oh et al. 2011]. These qualities are of particular importance in real-life systems since users are most likely to consider only a small set of top-N recommendations. It is therefore crucial to make sure that this set is as interesting and engaging as possible. In this article, we survey and analyze the most discussed objectives that relate to the quality of recommender systems beyond accuracy—diversity, serendipity, novelty, and coverage.

Before receiving attention in RS research, *diversity* and its relationship to accuracy were studied in information retrieval (IR) [Carbonell and Goldstein 1998] and, before that, in economics research. Markowitz [1952] introduced the Modern Portfolio Theory where investment is modeled as a tradeoff between risk and expected return. Maximizing the expected return results in higher investment risk, while diversification of stock portfolios reduces the risk. This idea has been adopted in IR [Carbonell and Goldstein 1998; Clarke et al. 2008; Wang and Zhu 2009; Agrawal et al. 2009], where it is argued that ranking retrieved items by only their predicted relevance (i.e., maximizing retrieval accuracy) increases the risk of producing results that do not satisfy users because the items tend to be too similar to each other. Conversely, diversifying the retrieval results reduces this risk by increasing the chance of introducing items the user will be interested in. In RS research, diversity is becoming an increasingly important topic, with a growing consensus that users are more satisfied with diverse recommendation lists, even if the diversity comes at a cost of some loss of accuracy [Ziegler et al. 2005; Shi et al. 2012; Vargas et al. 2014].

Serendipity is another objective that has received substantial attention in RS research. The term *serendipity*, referring to the process of "finding valuable or pleasant things that are not looked for,"[1] was coined in the 18th century [Van Andel 1994]. This objective is frequently mentioned in the IR and RS research literature [Toms 2000; André et al. 2009; Herlocker et al. 2004; Ge et al. 2010], where it is commonly agreed that serendipity consists of two components—surprise and relevance [Herlocker et al. 2004]. Until recently, however, few works provided formal definitions of metrics for measuring the serendipity of recommended items. This is not surprising, as the notion of an item being *surprising* or *unexpected* is difficult to define and measure.

Novelty is a recommendation quality that seems to be closely related to serendipity [McNee et al. 2006]. A novel recommended item is one that is previously unknown to the user. While the definitions may overlap [Zhang 2013], several authors distinguish novelty from serendipity. Herlocker et al. [2004] argued that an item that is novel to a user is not necessarily serendipitous for that user (it needs only to be unknown to the user), while a serendipitous item must be both novel and surprising; hence, the set of items that are serendipitous to a user is a subset of the set of items that are novel to that user. Adamopoulos and Tuzhilin [2014], on the other hand, defined a new objective (closely related to serendipity)—unexpectedness—and did not require an unexpected item to be novel. To better distinguish these two objectives, it is increasingly common to define the novelty of an item in a user-independent way, rather than the novelty of a recommended item to a target user. Typically, the novelty of an item is estimated by the inverse of its popularity (e.g., measured by the number of ratings it has received): items with low popularity are more likely to be new to target users [Celma 2009; Zhou et al. 2010]. By this definition, an item with high novelty will not necessarily be serendipitous for a user, and a serendipitous recommendation will not necessarily be novel.

[1] http://www.merriam-webster.com/dictionary/serendipity.

Coverage reflects the degree to which the generated recommendations cover the catalog of available items [Herlocker et al. 2004; Ge et al. 2010; Adomavicius and Kwon 2012]. Higher coverage may benefit both system users and business owners—exposing the users to a wider range of recommended items may increase their satisfaction with the system [Adomavicius and Kwon 2012] and also increase overall product sales [Anderson 2006]. In the literature, coverage is often linked to other beyond-accuracy objectives, particularly to novelty [Anderson 2006; Fleder and Hosanagar 2009; Adomavicius and Kwon 2012]. However, the relation between these objectives has not been extensively studied.

It is important to note that beyond-accuracy objectives may be pursued to a different extent in different recommendation scenarios, since the need for diversity, novelty, or serendipity may vary depending on the system's domain or user's needs. For instance, when recommending music, it is not always desirable to recommend unknown or surprising artists, as it may be important to include artists the user is familiar with but has not listened to in a while [Kapoor et al. 2015]. Indeed, in many domains, including a few familiar items among the recommendations may build trust in the system [Swearingen and Sinha 2001].

Moreover, the extent to which these objectives should be pursued may need to be adapted to each user's needs or preferences. For instance, when recommending movies, the level of diversity may be adapted to the user's range of tastes [Shi et al. 2012]. Likewise, the level of recommendation novelty may reflect the extent to which the user is interested in novel items [Oh et al. 2011]. While the adaptive aspect of beyond-accuracy objectives has not been extensively researched, in the following sections we highlight works that address this important problem.

In this article, we survey the definitions and optimization strategies for each of the objectives, and, using an empirical analysis, we investigate the relationships between them. Our work complements other surveys that cover various topics in RS research, such as recommendation algorithms [Ekstrand et al. 2011; Cacheda et al. 2011], side information in rating-based recommender systems [Shi et al. 2014], and evaluation metrics [Gunawardana and Shani 2009; Bellogín et al. 2011]. Recently, Castells et al. [2015] presented a survey closely related to ours. They reviewed different formulations of the diversity and novelty objectives found in the RS literature and analyzed the corresponding metrics. Compared to the work of Castells et al., we extend the analysis of beyond-accuracy objectives with experiments demonstrating how optimizing one beyond-accuracy criterion affects the other objectives. Thus, the contribution of our work is twofold: (1) we provide an extensive review of definitions and optimization techniques for the beyond-accuracy objectives, and (2) we conduct a number of experiments that demonstrate important insights into relationships between diversity, serendipity, novelty, and coverage. We hope that this work will become a useful reference for both researchers and practitioners working on beyond-accuracy objectives in recommender systems and will contribute to further growth of this research area.

Finally, we note that the terminology concerning beyond-accuracy objectives in the RS literature is not consistent. For instance, the term *diversity* is often used in reference to the system's ability to recommend different items to different users, or to the portion of the item catalog recommended across all users (i.e., coverage). In the ensuing sections, we cite existing works in the places where they best fit conceptually, regardless of the terminology used by the authors.

2. DIVERSITY

In this section, we first discuss the definition of diversity and the metrics proposed for measuring the diversity of recommendations. Subsequently, we review the techniques for increasing diversity.

2.1. Defining and Measuring Diversity

The notion of diversity in recommender systems originates from ideas in information retrieval research. In the IR literature, it has been acknowledged that the value of a retrieved document is influenced not just by the document's similarity to a query (its relevance), but also by its similarity to other documents retrieved with it [Carbonell and Goldstein 1998]. In information retrieval, the role of diversity is typically associated with possible ambiguity in a user's query—a search term *jaguar* may refer to the car, the animal, or the classic Fender guitar, for example [Clarke et al. 2008]. In the absence of disambiguating information, it is impossible to know which topic the user is interested in. Thus, ensuring that the list of retrieved documents covers a broad area of the information space increases the chance of satisfying the user's information need. This can be achieved by optimizing the diversity of the document list, which can be measured in terms of features (e.g., document types, information facts, topics) that the documents in the list possess [Carbonell and Goldstein 1998; Clarke et al. 2008; Wang and Zhu 2009; Agrawal et al. 2009].

In recommender systems research, Smyth and McClave [2001] suggested measuring the diversity of a recommendation list R ($|R| > 1$) as the average pairwise distance between items in the list:

$$Diversity(R) = \frac{\sum_{i \in R} \sum_{j \in R \setminus \{i\}} \text{dist}(i, j)}{|R|(|R| - 1)}. \quad (1)$$

Similarly, Ziegler et al. [2005] defined the "intra-list similarity" metric as the aggregate (rather than the average) pairwise similarity of items in the list, with higher scores denoting lower diversity of the list.

Measuring diversity as the average or aggregate dissimilarity of items in the recommendation list has been widely adopted in the RS literature. What often differs is the item distance function that is used ($\text{dist}(i, j)$ in Equation (1)). For instance, where items are represented by content descriptors, the distance between items has been measured using a taxonomy-based metric [Ziegler et al. 2005], the complement of Jaccard similarity [Vargas and Castells 2011], or the complement of cosine similarity on term vectors [Ekstrand et al. 2014]. Alternatively, where items are represented by rating vectors, item distance has been measured using Hamming distance [Kelly and Bridge 2006], the complement of Pearson correlation [Vargas and Castells 2011], or the complement of cosine similarity [Ribeiro et al. 2012].

Yu et al. [2009] suggested measuring item distance using the neighborhoods that are used for rating prediction in collaborative filtering (CF). In the case of item-based CF, each recommended item is represented by a neighborhood of items, while in the case of user-based CF, an item is represented by a neighborhood of users who rated the item. Item distance can then be computed as, for example, the complement of Jaccard or cosine similarity between the two items' neighborhoods.

Finally, item distance can also be obtained from the latent feature vectors in matrix factorization approaches [Vargas et al. 2011; Willemsen et al. 2011; Shi et al. 2012; Su et al. 2013].

Diversity metrics based on item dissimilarity were criticized by Vargas et al. [2014], who argued that the metrics fail to ensure that lists with high metric values will also be perceived by users as diverse. In domains where items can be described by sets of genres, Vargas et al. suggested using the genres for defining the diversity of an item list, arguing that genre diversity better corresponds to users' perception of diverse recommendations. They proposed three criteria that a genre-based diversity metric should capture—coverage, redundancy, and size awareness. In other words, a diversity metric value should reflect how well a list of items covers the genres a user is interested

in and how well genre redundancies are avoided. Moreover, it should be sensitive to the size of the recommendation list, since coverage and redundancy need to be treated differently for lists of different length.

Vargas et al. claimed that the optimal distribution of genres (in terms of diversity) is achieved when sampling items randomly. This idea is similar to the "diversity by proportionality" information retrieval approach by Dang and Croft [2012], who considered a list of retrieved documents most diverse when the number of documents covering each topic is proportional to the topic's popularity in the document corpus. Vargas et al. suggested a probabilistic model to measure genre diversity in a recommendation list. They proposed a "binomial diversity" metric that captures how closely the genre distribution in the item list matches the distribution that would be obtained by randomly sampling items from the dataset.

Since the balance between the diversity and accuracy of results is a widely discussed topic in information retrieval and recommender systems research, some works defined metrics that combine diversity and relevance. For instance, in IR research, Clarke et al. [2008] described α-*nDCG*—a diversity-aware ranking measure, where the score of retrieved documents is penalized if they share features with documents ranked higher in the list. In RS research, Vargas and Castells [2011] proposed a framework in which the diversity of a recommendation list can be computed with a relevance and ranking discount. The authors argued that irrelevant recommendations add little to the perceived diversity of a recommender, making it necessary to weight the diversity score with the items' relevance.

Other diversity definitions, not referring to the quality of a single recommendation list, can also be found in the RS literature. For instance, Lathia et al. [2010] analyzed how recommendations generated for the same user change over time. They defined "temporal diversity" as the normalized set theoretic difference between top-N recommendation lists received by the same user at two different time points. Averaging the values across all users gives an estimate of the system's ability to provide users with diverse recommendations over time. Diversity has also been defined from a system-centric perspective, for example, as the average pairwise distance between recommendation lists generated for different users [Zhou et al. 2010; Liu et al. 2012]. These definitions do not fit the view of diversity that we adopt in this article. Therefore, in the next section, we focus on works that optimize the diversity of an individual user's recommendation list.

2.2. Increasing Diversity

Most diversification techniques in the RS (and also IR) literature are based on reranking the result lists generated by existing recommendation (and retrieval) algorithms to increase their diversity while maintaining relevance. Another group of approaches includes works that define new models for diversity-oriented recommendation. We discuss both groups of techniques in detail.

2.2.1. Recommendation Reranking for Diversity. The reranking diversification approaches produce a list of recommended items R of size N from a larger set of candidate recommendations C ($|C| > N$). The candidates C are generated by an existing recommendation algorithm (e.g., user-based collaborative filtering), and hence have been chosen for their relevance. Reranking typically follows a greedy strategy: at each iteration, the item in C that maximizes an objective function is moved from C to result list R. The objective function is defined as a combination of an item's relevance and its relative diversity with respect to items already in the result list R. The greedy reranking is illustrated in Algorithm 1.

ALGORITHM 1: The Greedy Reranking Algorithm. (We Use "++" to Denote List Concatenation and "\" to Denote Set Difference.)

Data: N; a set of candidate items C, s.t. $|C| > N$
Result: result list R, s.t. $|R| = N$

$R \longleftarrow [\,]$;
while $|R| < N$ **do**
  $i \longleftarrow \arg\max_{i \in C} f_{obj}(i, R)$;
  $R \longleftarrow R ++ [i]$;
  $C \longleftarrow C \setminus \{i\}$;
end
return R;

One of the early diversification techniques to use greedy reranking is the Maximal Marginal Relevance (MMR) approach proposed by Carbonell and Goldstein [1998] in the IR literature. The MMR approach defined the objective function f_{obj} as a linear combination of the item's relevance and the negative of its maximum similarity to items already in the result list.

The greedy reranking technique has been adopted by a number of recommendation approaches [Smyth and McClave 2001; Ziegler et al. 2005; Kelly and Bridge 2006], which defined the objective reranking function as a linear combination of the item's relevance and its average distance to items already in the result list:

$$f_{obj}(i, R) = \alpha \cdot rel(i) + (1 - \alpha) \cdot \frac{1}{|R|} \sum_{j \in R} \text{dist}(i, j). \tag{2}$$

In the equation, $rel(i)$ denotes the item's relevance and parameter α controls the trade-off between the influence of relevance and diversity in the reranking procedure. Similarly to the diversity metric (see Section 2.1, Equation (1)), the distance between two items $\text{dist}(i, j)$ can be computed using a variety of approaches.

Smyth and McClave [2001] applied the technique in a case-based recommender where, given a user's query, the database of cases is searched to retrieve the most relevant cases. In this setting, $rel(i)$ represents the similarity between the user's query and a case, while $\text{dist}(i, j)$ is the complement of the similarity between two cases.

Ziegler et al. [2005] applied the reranking technique for book recommendation, where the list of recommendations is generated based on the user's rating profile (i.e., using a CF algorithm). The authors defined $rel(i)$ as the item's relevance predicted by the recommender, and $\text{dist}(i, j)$ as the distance between two items, this being obtained from a genre taxonomy-based metric. Ziegler et al. were also the first to conduct a user study analyzing the impact of diversification on user satisfaction with the recommendation list (see Section 6).

Kelly and Bridge [2006] applied the greedy reranking strategy in a conversational CF recommender, where recommendations are presented to a user through a series of interaction cycles—after receiving a set of recommendations, the user provides feedback, which influences the next set of recommendations. The dialog is repeated until the user is satisfied with the provided recommendations. The authors proposed to diversify the set of recommendations at each interaction cycle, with $rel(i)$ as the predicted item's relevance and $\text{dist}(i, j)$ computed as the normalized Hamming distance of the two items' binary rating vectors.

The setting of conversational recommendations poses additional challenges for result diversification. McGinty and Smyth [2003] pointed out that the level of diversity can be varied in different recommendation cycles. The authors described a system where at

each recommendation cycle, the user selects the best-quality recommendation, which is used as a query for the next cycle. The selected recommendation is carried over to the next cycle and displayed alongside the newly generated recommendations. If the user selects the carry-over item again, the system concludes that no progress toward the user's goal has been made and injects more diversity in the next cycle. If, however, the user selects a recommendation different from the carry-over item, the system assumes positive progress has been made and generates more similar recommendations for the next cycle.

Recent work on recommendation reranking for diversity has focused on designing more advanced objective functions that combine item relevance and diversity. For example, Vargas et al. [2011] suggested applying diversification techniques and metrics from IR research to the recommender systems domain. They adopted the objective function from the IA-Select approach proposed by Agrawal et al. [2009]. IA-Select is a probabilistic model similar to the greedy reranking approach (see Algorithm 1), which assumes a feature space of information topics, such that both documents and user queries can be represented with a distribution over the feature space. The model defines an objective reranking function that considers both document relevance and topic distribution, thus avoiding topic redundancy in the result list. To adapt the IA-Select model to a recommender setting, Vargas et al. suggested replacing the topic feature space with either item labels (e.g., genres) or the latent item feature space (extracted using a matrix factorization approach).

Other recent reranking work by Vargas et al. [2014] proposed a "binomial diversity" metric to measure genre diversity in a recommendation list (see previous section). The authors used greedy reranking with an objective function that combines item relevance with its relative binomial diversity (i.e., the difference in the binomial diversity of the result set before and after adding the item).

Barraza-Urbina et al. [2015] proposed another formulation of the objective function for the greedy reranking strategy. They suggested explicitly controlling the level to which diversification promotes items that are dissimilar to the user's profile items. This was achieved by multiplying the diversity component ($\frac{1}{|R|}\sum_{j\in R}\text{dist}(i,j)$ in Equation (2)) by a weighted combination of exploration and exploitation scores for item i: $\beta \cdot xploit(i) + (1-\beta)\cdot xplore(i)$. The exploitation score $xploit(i)$ measures the probability that items in the user's profile that are similar to i have been highly rated by the user, and the exploration score $xplore(i)$ captures the item's average dissimilarity from the items in the user's profile. The β parameter thus controls the balance between a more "safe" diversification, picking diverse items that are within the user's known taste range, and a more "explorative" diversification, promoting serendipitous items (see Section 3).

There are also reranking techniques that do not use a greedy reranking strategy. Typically, they rely on solving optimization problems to find the optimal ranking for a list of candidate recommendations. For instance, Zhang and Hurley [2008] used an item-based CF approach to compute an item-to-item similarity matrix and then solved a number of optimization problems to find the set of recommended items that maximizes the diversity while maintaining a certain level of accuracy. The authors used the term "item novelty" to denote the amount of additional diversity that an item brings to the recommendation set.

Jambor and Wang [2010] proposed a generic constrained optimization framework that supports multiple beyond-accuracy objectives. The authors suggested predicting item relevance scores using existing recommendation techniques and weighting them with utility weights. Item relevance is specified as the main objective in the framework, and additional constraints can be defined for the utility weights to optimize for diversity or novelty (Section 4).

Ribeiro et al. [2012] proposed an optimization approach similar to recommendation reranking—rather than reordering recommendations generated by a single algorithm, they used the relevance scores predicted by different algorithms in a weighted combination to determine the final item utility score. They focused on three objectives—accuracy, diversity, and novelty—and hypothesized that a hybrid combination of different algorithms can provide a better balance of the objectives. The baseline algorithms used within the weighted combination included three variants of the matrix factorization approach, user-based and item-based nearest-neighbor approaches, a popularity-based approach, and simple content-based and demographic-based nearest-neighbor methods. The weights for each algorithm were learned using a genetic algorithm. Since the three objectives are potentially conflicting, the approach selected the weight combinations that are optimal on the Pareto frontier, that is, the solutions where none of the three objectives can be improved without hurting the other two. Diversity and novelty were defined using the rank-aware metrics proposed by Vargas and Castells [2011].

2.2.2. Diversity Modeling. The reranking techniques described in the previous section treat recommendation algorithms as a "black box." They work by postprocessing lists of items that are generated by the recommendation algorithms. An obvious advantage of the reranking techniques is their ease of deployment in existing recommender systems, where a diversification component may be incorporated alongside existing recommendation algorithms and the level of diversification can be explicitly controlled. However, there is a growing body of research that addresses the diversification problem by defining new recommendation algorithms that directly optimize for diversity when generating recommendations. These approaches mostly extend matrix factorization techniques, which have become the state-of-the-art recommendation methods in recent years.

For instance, Shi et al. [2012] combined matrix factorization with the portfolio theory from IR proposed by Wang and Zhu [2009] (whose work in turn was inspired by the Modern Portfolio Theory from economics [Markowitz 1952]). The IR portfolio theory considers the predicted document relevance as an uncertain outcome whose expected value may be over- or underestimated (due to query ambiguity, incomplete user profile, imperfect retrieval algorithm, etc.). Given the uncertainty of document retrieval, a probabilistic model is used to represent the expected overall relevance of the retrieved document list and its variance. The variance of the list represents the likelihood that the relevance of the documents was estimated incorrectly and is computed using the covariance of document relevance scores for each document pair in the list. The covariance of document relevance scores can be approximated by term co-occurrence in the documents.

The basic idea of the portfolio theory is to minimize the risk of generating an item list with low relevance for the user. This is achieved by maximizing the expected relevance and minimizing the variance of the result list. Shi et al. adapted this idea to a recommendation setting and defined an objective function that balances the predicted relevance and variance of the recommendation list. Differently from the IR work, where variance was approximated by term co-occurrence in documents, Shi et al. used latent factor vectors (obtained from the matrix factorization approach) to model the variance of recommendations.

An important aspect of the approach of Shi et al. is adapting the level of diversification to the user's scope of tastes. They showed that the latent factors of a user who rates diverse items have higher variance compared to a user who rates similar items. This is reflected in the proposed model. Therefore, a user who tends to rate similar items (e.g., only science fiction movies) will get less diversification compared to a user who rates diverse items (e.g., movies from different genres).

Hurley [2013] presented a modification of the pairwise learning-to-rank approach for implicit feedback datasets. The original pairwise ranking model learns the user and item factors by minimizing an objective function defined on the difference between the predicted and original ranking for item pairs. In the modified diversity-aware version of the model, Hurley proposed including item dissimilarity in the objective function. Although the learning model is not sensitive to the size of recommendation list N (when generating top-N recommendations), an initial evaluation of the approach using the "intralist distance" metric [Smyth and McClave 2001; Ziegler et al. 2005] (Equation (1)) showed promising results.

Su et al. [2013] proposed a pairwise learning-to-rank model that works at the item set level (rather than the individual item level). The training data is constructed by creating pairs of item sets. The model is trained by comparing each pair of item sets using both relevance and diversity criteria. Diversity of a set is included in the model through a "diversity bias" component, defined as the aggregate similarity of all item pairs in the set. The similarity of an item pair is computed as the product of the two items' latent factor vectors.

3. SERENDIPITY

3.1. Defining and Measuring Serendipity

Defining serendipity largely relies on the definition of its core component—*surprise*. In the cognitive science literature, surprise has been linked to events that are different from one's expectations [Meyer et al. 1997] or are difficult to explain [Foster and Keane 2013]. Such definitions are not trivial to operationalize in the information retrieval or recommender systems domain.

The first studies that recognized the importance of facilitating serendipity in information systems were reported in the IR literature [Toms 2000]. Rather than providing formal definitions of serendipity, early works analyzed the process of serendipitous information discovery and the paradox of designing for unexpected results [Foster and Ford 2003; McBirnie 2008]. In the RS literature, Herlocker et al. [2004] informally defined a serendipitous recommendation as one that helps the user find a "surprisingly interesting item he might not have otherwise discovered."

In the RS literature, approaches designed to increase serendipity rely on various heuristics to generate more surprising recommendations. For instance, an item can be considered serendipitous if a classifier is uncertain about its relevance for the user [Iaquinta et al. 2008], if the item is different from the user's profile [Adamopoulos and Tuzhilin 2014], if the item is connected to a distinct area in a user-item graph [Onuma et al. 2009; Nakatsuji et al. 2010; Zhang et al. 2012], or if the item possesses a mixture of two input items' features [Oku and Hattori 2011] (see next section for details).

When using offline experiments to evaluate the quality of results produced by these ad hoc approaches, a common practice among the authors is to compare the generated recommendations with recommendations produced by a primitive baseline system (i.e., one that is not optimized for serendipity). This approach to measuring serendipity was first proposed by Murakami et al. [2008], who argued that a primitive method produces easily predictable items, while the goal of a serendipitous recommender is to suggest items that are difficult to predict. This idea was later adopted by Ge et al. [2010], who proposed a formulation of serendipity that combines this notion of unexpectedness with item relevance:

$$Serendipity(R, u) = \frac{|R_{unexp} \cap R_{useful}|}{|R|}, \tag{3}$$

where R is the set of recommendations generated for user u, R_{unexp} is the subset of items in R that are *unexpected* for the user u, and R_{useful} is the subset of items in R that

are *useful* for the user. Following the idea of Murakami et al., the set of unexpected recommendations is obtained by subtracting from R items that are recommended by a primitive prediction model PM for user u: $R_{unexp} = R \setminus PM_u$. The usefulness of recommendations may be judged by the user or, in an offline setting, approximated by the user's ratings for the items [Adamopoulos and Tuzhilin 2014]. A limitation of this comparative approach to serendipity measurement is its sensitivity to the choice of the primitive baseline system.

Recently, Adamopoulos and Tuzhilin [2014] suggested another way to measure the unexpectedness of recommendations. The authors defined R_{unexp} as $R \setminus E_u$, where E_u is the set of *expected* recommendations for a user u, which contains items rated by the user and items that are similar to the rated ones (in terms of content similarity). Note that contrary to the original (informal) definition of serendipity by Herlocker et al. [2004], metrics like these—based on item unexpectedness—do not require serendipitous items to be novel to the user, but only relevant and different from the user's expectations.

The idea of measuring an item's unexpectedness as its distance from a set of expected items has been exploited by a few previous works. Nakatsuji et al. [2010] proposed an approach based on a taxonomy of genres and defined what they called "item novelty" as the smallest distance (in the taxonomy) from the item's genre to the genre of items previously accessed by user. Vargas and Castells [2011], in their framework for measuring diversity and novelty, defined what they called a "personalized novelty" metric based on computing an item's average distance from the user's profile items.

In our experiments (see Section 7), we adopt the idea of measuring an item's unexpectedness (or surprise) as its distance from the set of expected items. Furthermore, we follow the idea of Nakatsuji et al. to measure an item's surprise as the *minimum* distance from the user's profile items and we hypothesize that, by contrast, *averaging* the distances between items results in a loss of information, particularly for users with diverse profiles [Kaminskas and Bridge 2014].

3.2. Increasing Serendipity

The first attempts to increase the serendipity of retrieved results were reported in the IR literature. For example, Campos and de Figueiredo [2001] designed a software agent to support serendipitous information discovery through web crawling. André et al. [2009] suggested viewing serendipity as a combination of chance discovery and usefulness of the discovered information. They provided guidelines to design information systems with better support for both components of serendipity: supporting chance encounters and enhancing the user's ability to recognize serendipitous content.

In the RS literature, Iaquinta et al. [2008] were among the first to introduce serendipity in a recommender system. They described a content-based recommender with items represented by text descriptions. A supervised learning method was used to predict the probability that an unseen item was either relevant or nonrelevant to the user. Items for which the classification outcome was uncertain (i.e., where the absolute difference between the two probabilities was large) were considered as potentially serendipitous and were included in the recommendations.

Onuma et al. [2009] designed a system that uses a graph-based algorithm for supporting surprising recommendations. The authors introduced the idea of computing a "bridging score" for item nodes in the user-item bipartite graph. Nodes connecting separate interconnected areas in the graph receive high bridging scores as they bridge different subspaces in the item information space. The bridging score may be combined with an item relevance score when generating recommendations. Another graph-based approach was proposed by Nakatsuji et al. [2010]. They applied a Random Walk algorithm on a user similarity graph to identify users that are related (but not too

similar) to the target user, arguing that such users provide a good source of surprising recommendations.

Oku and Hattori [2011] presented a system that induced possibly serendipitous recommendations by selecting items whose content is a mixture of the content features of two items from the user's profile. Zhang et al. [2012] presented a music recommender for Last.fm artists that uses a generative Latent Dirichlet Allocation (LDA) model to build latent clusters of Last.fm users and to represent artists by a distribution over these clusters. Representing artists as LDA vectors gives a way of computing a similarity score between any artist and the artists in a user's listening profile. Moreover, the vector representation allows artists to be clustered. The recommender generates serendipitous recommendations by promoting artists that are outside of the user's "musical bubbles" (clusters of liked artists).

Finally, Adamopoulos and Tuzhilin [2014] presented an approach to recommend serendipitous items based on how distant they are from the set of items expected by the user. The authors defined an item utility function as a linear combination of the item's relevance score (predicted by a standard recommendation algorithm, e.g., a collaborative filtering approach) and its *unexpectedness* (computed as a distance between the item and a set of expected items). The set of expected items includes items rated by the target user and items similar to the rated ones in terms of content (e.g., in a movie domain, movies produced by the same director and belonging to the same genre). For computing the distance between an item and a set of items, the authors suggested averaging the individual distance values or computing the centroid of the set and measuring the target item's distance from the centroid. Both rating- and content-based distance metrics were evaluated. Given the target user and a set of the user's expected items, the proposed recommendation approach computes the utility score for each candidate item and recommends those with the highest utility values.

4. NOVELTY

Novelty is closely related to serendipity, discussed in the previous section. Here, we first discuss the relation between these two objectives and motivate our choice of novelty definition. Subsequently, we discuss research that addressed novelty optimization in recommender systems.

4.1. Defining and Measuring Novelty

Similarly to other objectives discussed in this work, the definition of novelty in the RS literature is inspired by IR research. Baeza-Yates and Ribeiro-Neto [1999] were among the first to discuss novelty as an important quality in information retrieval. They defined the novelty of a set of retrieved documents as the fraction of relevant documents that are unknown to the user. Another view on novelty was offered by Zhang et al. [2002], who considered the novelty of a single retrieved document as the opposite of its redundancy. They proposed a number of redundancy metrics based on the distance between the document and documents previously seen by the user.

The aforementioned views of novelty are both related to how novelty is commonly perceived—"the quality or state of being new, different, and interesting."[2] Definitions of novelty in the RS literature typically focus on two aspects of novelty—an item being unknown to the user and an item being different from what the user has seen before. Some works focused only on the latter aspect and proposed novelty metrics that measure an item's distance from the user's profile (i.e., previously seen items) [Yang and Li 2005; Nakatsuji et al. 2010]. Vargas and Castells [2011] proposed different variants of a novelty metric that support both the *unknown* and the *different* aspects

[2]http://www.merriam-webster.com/dictionary/novelty.

of novelty. Zhang [2013] identified three qualities of a novel recommendation: being unknown to the user, being relevant to the user, and being dissimilar to items in the user's profile.

We note that the quality of an item being different from the user's profile is closely related to the *surprise* of recommendations, which we identify as a core component of serendipity (see Section 3). As discussed in Section 1, novelty and serendipity are closely related and their definitions in the literature often overlap. A distinction between the two objectives was proposed by Herlocker et al. [2004], who argued that a novel item does not have to be surprising, but only unknown to the user. Kapoor et al. [2015] extended the definition of novel items to include those that are known but forgotten by the user (i.e., items the user has not accessed in a while). A "temporal novelty" formulation like this one is only applicable to domains with frequent repeated item consumption, for example, music recommendation. To better structure the discussion of serendipity and novelty, in this section we follow the definition of Herlocker et al. [2004] and view novel recommendations only as those that are unknown to the user.

The quality of an item being unknown is not trivial to define formally. While a typical recommender provides a user with suggestions for *unrated* items, an absent rating does not necessarily imply an unknown item—a user rarely provides ratings for all known items. Therefore, without acquiring the user's feedback (e.g., through a user study), it is impossible to know if an unrated item is truly novel. Hijikata et al. [2009] proposed a CF system where two types of rating profiles were created for each user—the traditional rating profile, containing the item preferences, and the "acquaintance profile," containing binary ratings of item familiarity (i.e., "known/unknown"). The authors suggested a number of hybrid CF algorithms exploiting the two types of profiles to generate both unknown and accurate recommendations. The approach, although explicitly addressing the issue of item novelty, doubles the cognitive load of user profile construction as the users need to provide both types of ratings. More commonly, an item's novelty is approximated using its popularity among users of the recommender system—the less popular an item is, the more likely it is to be unknown to the user.

Although an item's unpopularity is not always a good indication of it being unknown—a user familiar with one rare item is likely to know similar rare items [Celma 2009]—it provides a cheap approximation for measuring novelty offline, without conducting costly user studies. Item popularity has been estimated using rating variance [Jambor and Wang 2010] in the dataset or using external sources of information, such as box office earnings for movies [Oh et al. 2011]. However, the most common approach is based on the number of ratings an item has received from users.

Formally, then, novelty is typically defined as the complement of the item's popularity in the dataset: $1 - p(i)$, where $p(i) = \frac{|\{u \in U, r_{ui} \neq \emptyset\}|}{|U|}$ is the fraction of users who rated item i. A slight variation is to define novelty as the negative of the log of the ratio: $-\log p(i)$. This formulation is called the *self-information* of an item i [Zhou et al. 2010; Vargas and Castells 2011] and, compared to the simple complement of popularity, gives more importance to very rare items.

In order to evaluate recommendation techniques with respect to novelty, the novelty of individual recommendations is aggregated into a single score for a list of recommendations R:

$$Novelty(R) = \frac{\sum_{i \in R} -\log_2 p(i)}{|R|}. \tag{4}$$

Given the previous definition of novelty, novel items are identified with the "long tail" items, that is, the part of the item catalog seen (rated or purchased) by a small part of the user community [Anderson 2006]. A detailed analysis of the long-tail phenomenon and its influence on recommendation novelty was given by Celma [2009].

Celma analyzed the long-tail item distribution and its relation to item similarity in a music recommender. The recommender system was modeled as a fully connected graph with nodes representing items; edges were weighted by similarities between items. Two versions of the recommender were analyzed—an item-based CF approach and a content-based (CB) approach. The item similarities on the edges were rating based for the CF system and content based for the CB system. Celma compared the long-tail distribution of item popularity with the item similarity graph and showed that, in the CF system, popular items tend to form highly interconnected clusters in the graph, meaning that the long-tail (i.e., novel) items are difficult to reach and therefore difficult to recommend to users. Conversely, in the CB system, item connections in the graph are independent of their popularity, therefore making CB recommendations more novelty oriented.

4.2. Increasing Novelty

Based on the definition of novelty adopted in this work (i.e., based on item popularity), in this section, we focus on works that increase recommendation novelty by promoting rare items (also known as the "long tail" items).

One of the early efforts to analyze the long-tail phenomenon in recommender systems was presented by Park and Tuzhilin [2008]. Although not directly related to novelty optimization, this work dealt with improving rating prediction accuracy for the long-tail items. They observed that when using rating-based prediction algorithms, prediction accuracy for rare items is lower than for popular items (caused by the smaller number of ratings on which the prediction is based). Their suggested solution for improving the prediction accuracy was based on clustering the long-tail items and creating joint rating profiles for the clusters. Then, for a given long-tail item, rating prediction could be made using all ratings in its cluster. Experiments with the MovieLens dataset showed reduced error rates using the proposed approach. However, the proposed technique did not guarantee promotion of the long-tail items into users' top-N recommendation lists.

Ishikawa et al. [2008] addressed the long-tail phenomenon in the context of recommending knowledge resources (web pages) in an information portal. They proposed an approach based on innovation diffusion theory, claiming that new information spreads among users according to observable patterns, with the first users to access a resource playing the role of "innovators." The proposed algorithm therefore requires a "seed" long-tail item and exploits the users who first "discovered" it as a source of novel recommendations (by recommending other items that were accessed by these users). Experiments with the portal's log data showed 10 to be the optimal number of "innovators." Since the approach was designed within a very specific domain, it would be interesting to evaluate it in more common recommendation domains, such as movies or music.

Zhou et al. [2010] exploited item popularity information to increase both novelty (measured as the *self-information* of recommended items, see Equation (4)) and interuser diversity (average pairwise distance between recommendation lists generated for different users). The authors proposed an algorithm based on weight spreading in a bipartite user-item graph. The algorithm works by assigning weights to items rated by the target user and then equally distributing the weight of each item to other users who rated it. The weight of each user is then distributed among his or her rated items. This weight spreading procedure favors item nodes with few graph links (i.e., rare items), resulting in novel recommendations. In similar research, Liu et al. [2012] described a graph-based algorithm with weight spreading and showed that assigning more weight to users with small profiles enhances both interuser diversity and the novelty of recommendations.

Oh et al. [2011] followed the idea that novelty is related to both popularity and the interuser diversity of recommendations. They worked with the MovieLens dataset and measured item popularity using movie box office earnings (applying a log scale to smooth the effects of the power-law distribution). The work demonstrated that a state-of-the-art method for item-based collaborative filtering and a novelty-optimized recommender (the Tangent system [Onuma et al. 2009], which we discussed in the context of serendipity, see Section 3) both perform poorly in terms of popularity and interuser diversity; that is, their generated recommendations are clustered around popular items. Oh et al. argued that the users' preferences for popular items should partly determine the recommendations. From the users' rating histories, they identified different types of "personal popularity tendency," that is, different levels of user interest in rare/popular items. Their proposed novelty optimization approach reranks recommendations by penalizing items that do not fit the user's popularity tendency.

Another graph-based approach was proposed by Shi [2013], who defined a cost flow model for a bipartite user-item graph. The model is based on assigning a transition cost for each edge between a user node and an item node. Given a target user and candidate items, the cost to reach the target user node from a candidate item node can be computed by propagating the cost score through the edges. Items that obtain the lowest-cost scores are then recommended to the user. Shi proposed different strategies for defining the edge transition costs, including the "long tail" strategy that sets the costs to be proportional to the popularity of item nodes they connect, thus promoting rare items.

We observe that works addressing the long-tail recommendation problem often measure algorithm performance not only in terms of novelty but also in terms of interuser diversity [Zhou et al. 2010; Liu et al. 2012] and coverage [Shi 2013]. This indicates that novelty is closely related to these system-level objectives. Interuser diversity measures the difference between recommendations across different users, while coverage measures how well the recommender covers the available item catalog. In this article, we do not discuss the interuser diversity but focus on the more popular system-level objective—coverage.

5. COVERAGE

Unlike the beyond-accuracy objectives that we have discussed so far, coverage is not defined at the level of an individual user, but rather at the level of the system.

There are two general approaches to measuring recommendation coverage—"user coverage," which measures the degree to which the system covers its users (e.g., the ratio of users for which a recommender is able to deliver recommendations [Bellogín et al. 2013]), and "item coverage," which measures the degree to which recommendations cover the set of available items (i.e., the item catalog). Since the latter formulation is more commonly found in the RS literature [Herlocker et al. 2004; Ge et al. 2010; Adomavicius and Kwon 2012] and to be consistent with the other discussed beyond-accuracy objectives (which are item properties rather than user properties), in this work we focus on "item coverage" and henceforth refer to it simply by the term *coverage*.

Since measures of coverage show how well the system's recommendations cover the catalog of available items, higher coverage means exposing the users to a wider range of recommended items, which may both increase the users' satisfaction with the system (e.g., by recommending items the user would not otherwise discover) [Adomavicius and Kwon 2012] and benefit business owners (increasing the sales of the long-tail items). For instance, Anderson [2006] argued that aggregate sales of the long-tail products may match (or even outnumber) the sales of the top-selling products.

5.1. Defining and Measuring Coverage

As with other beyond-accuracy objectives, the terminology used to identify the coverage objective varies across different works. Coverage has been referred to as "aggregate diversity" [Adomavicius and Kwon 2011, 2012], "sales diversity" [Vargas and Castells 2014], or simply "diversity" [Shi 2013]. To avoid confusion with the diversity objective, which is applicable to a single user's list of recommendations (see Section 2), we use the term *coverage* and adopt its most widespread definition—the fraction of items that appear in the users' recommendation lists:

$$Coverage = \frac{|\cup_{u \in U} R_u|}{|I|}, \tag{5}$$

where R_u is the set of all recommendations generated for user u, U is the set of all users of the system, and I is the item catalog.

Herlocker et al. [2004] distinguished between two forms of item coverage—"prediction coverage," which captures the ratio of items for which predictions can be made by the recommendation algorithm, and "catalog coverage," which captures the ratio of items that effectively appear in the recommendation lists presented to users of the system. The metric adopted in our work (Equation (5)) corresponds to catalog coverage and we do not discuss prediction coverage separately.

A few works have suggested taking recommendation relevance into account when measuring coverage, that is, measuring the fraction of *relevant* items that are recommended to all users [Herlocker et al. 2004; Bellogín et al. 2013]. Such definitions require that we know all the potentially relevant items for each user. Ge et al. [2010] proposed a more general definition where each item contributing to the coverage score is weighted by its usefulness. The authors suggested that the usefulness weights may be computed using item relevance, novelty, or serendipity scores.

All the variants of the coverage metric discussed previously share a common feature—since all recommended items are aggregated into a single score, an item recommended once contributes to the coverage score the same amount as an item recommended a thousand times. An alternative definition, which overcomes this limitation, was proposed by Fleder and Hosanagar [2009]. They used the *Gini coefficient*, ranging in [0, 1], to measure the distribution of recommendations across all users. High values of the coefficient mean that recommendations are concentrated around a few frequently recommended items (i.e., there is a high *concentration bias* [Jannach et al. 2015b]), while lower values signify a more uniform distribution of recommendations. Vargas and Castells [2014] adopted the complement of the Gini coefficient so that higher values of the metric correspond to better (more uniform) catalog coverage.

Shani and Gunawardana [2011] identified *Shannon entropy* as another option for measuring the distribution of recommendations across users. For an item catalog of size n, the Shannon entropy metric ranges between 0 (when the same single item is always recommended) and $\log n$ (when all items are recommended equally often).

Despite being a system-level objective, coverage is related to other objectives discussed in this article, particularly to novelty. As discussed in Section 4, novelty is typically measured as the complement of item popularity. Highly novel recommendations are therefore those belonging to the long tail of the item popularity distribution. Intuitively, a high coverage of the item catalog requires recommending the long-tail items to users, which corresponds to a high average novelty (see Equation (4)).

However, the relation between coverage and novelty is not always straightforward, as has been shown by Jannach et al. [2013]. The authors compared a number of state-of-the-art recommendation algorithms in terms of coverage and the tendency to focus on popular items. In experiments performed on the MovieLens dataset, a

learning-to-rank algorithm achieved a high level of coverage but also suffered from popularity bias (i.e., having higher average popularity of recommended items compared to other algorithms).

Coverage has also been discussed in relation to diversity. Adomavicius and Kwon [2012] discussed the difference between diversity (which they call "individual diversity," i.e., the diversity of a recommendation list presented to a single user) and coverage (which they call "aggregate diversity," i.e., the range of items recommended across all system users). They argued that high diversity does not imply high coverage. For instance, if different users are recommended the same diverse set of items, the average diversity of the system will be high, but the coverage will remain low. Similarly, Fleder and Hosanagar [2009] discussed a scenario where (during the evolution of an e-commerce recommender) the system helps users discover new items, but the Gini coefficient increases (i.e., the coverage deteriorates). This happens if the users are exposed to new items, but these are the same items other users have seen before.

Ge et al. [2010] briefly discussed the relation between coverage and serendipity. They argued that high serendipity implies high coverage, but an increase in coverage will not necessarily improve serendipity. The authors, however, offered no experiments to support this hypothesis.

In summary, the close relation between coverage and other beyond-accuracy objectives has been recognized in the literature. However, there is a lack of experiments that study the relationships between the different metrics. We contribute to the analysis of this research problem with experiments presented in Section 7.

5.2. Increasing Coverage

As discussed earlier, coverage can be linked to the novelty of recommendations. This is reflected in work that addresses the coverage optimization problem, with most approaches relying on reducing the popularity bias of recommendations (i.e., increasing the number of long-tail items recommended to users). Consequently, the approaches discussed in this section overlap with those described in Section 4.2. To avoid repetition, here we discuss works that explicitly target increasing the coverage of recommendations.

Adomavicius and Kwon [2011] modeled the item-to-user recommendations (precomputed using a standard recommendation algorithm) as a graph, where an edge connects an item to a user only if the item is predicted as relevant to that user (a prediction threshold may be used when constructing the graph). They then solve the maximum flow problem on the constructed graph; that is, they find the maximum number of edges that connect users and items, such that each user is connected to no more than N items. The solution of this problem results in each user being assigned up to N recommendations with the maximum coverage of the available items.

Another work by Adomavicius and Kwon [2012] described a coverage optimization approach based on reranking the recommendation list to promote the long-tail items. The approach reranks items whose predicted rating is above a certain threshold by their popularity (ranking rare items higher). The threshold parameter guarantees a certain level of accuracy in the list and can be varied for a tradeoff between accuracy and coverage. Offline evaluation with the MovieLens dataset showed the approach to improve coverage with a minimal loss in recommendation accuracy.

Vargas and Castells [2014] proposed an approach to increase coverage by reducing the popularity bias in nearest-neighbor CF algorithms. They suggested inverting the neighbor selection process in user-based or item-based CF: for instance, in the case of item-based nearest-neighbor CF, instead of selecting the top-k most similar neighbors for an item, they suggested constructing the neighborhood by selecting those items in whose neighborhoods the target item appears. (A similar inversion can be used for

user-based neighborhood techniques.) The authors demonstrated that such inversion results in a reduced popularity bias, since all items appear in the same number of the newly defined neighborhoods and therefore rating prediction is less influenced by the popular items. Offline experiments with Netflix data and the Million Song Dataset showed that the inverted item-based CF approach outperformed the standard item-based technique in both accuracy and coverage (measured as the complement of Gini coefficient). The inverted user-based approach showed better coverage results but did not outperform the standard user-based technique in terms of accuracy.

6. MEASURING THE USER'S PERCEPTION OF BEYOND-ACCURACY OBJECTIVES

Any evaluation of recommendation diversity, serendipity, or novelty not involving user feedback is limited in terms of the reliability of the findings. For instance, without asking the end-user of the system, it is not evident that an item that was shown to be serendipitous by some metric will be perceived as such by the user. Likewise, it is not evident that a diversification algorithm that considers pairwise dissimilarity of recommended items will produce recommendation lists that users perceive as diverse. However, despite the obvious need for user feedback on the beyond-accuracy qualities of recommendations, few research works in this area include user studies, the majority of results being obtained from offline experiments. This can be explained by the many challenges in designing and conducting such studies: recruiting a sufficiently large number of participants, correctly formulating survey questions, avoiding judgment biases, and so forth. Moreover, when recommending items that take a long time to consume (e.g., movies), relevance, diversity, or serendipity judgments are difficult to obtain for items that are unknown to users.

In this section, we provide an overview of the limited number of works that do rely on user studies when analyzing beyond-accuracy objectives. We split such works into two general categories: (1) research studies where beyond-accuracy objectives are evaluated as a part of larger multicriteria experiments that measure relationships between the different recommendation qualities perceived by users (e.g., the impact of perceived novelty on diversity perception) [Pu et al. 2011; Knijnenburg et al. 2012; Ekstrand et al. 2014], and (2) works that analyze the impact of the proposed algorithms (or user interface modifications) on specific beyond-accuracy objectives [Ziegler et al. 2005; Ge et al. 2012; Hu and Pu 2011].

Of the beyond-accuracy objectives discussed in our work, *diversity* and *novelty* are the ones that are most frequently investigated in user studies. *Serendipity* has been reported to be difficult to explain to users [Said et al. 2013] or has been left out of the studies as being too similar to novelty [Pu et al. 2011]. *Coverage* is not measured in user studies since it is not directly related to individual user experiences. Besides the perceived accuracy, diversity, and novelty, user studies often also measure user *satisfaction* with the system. Although satisfaction is a concept easily understood by users, we consider it as a higher-level quality that can be influenced by many perceived qualities (relevance, diversity, novelty, serendipity) and therefore do not analyze it in this article.

6.1. Beyond-Accuracy Objectives in Multicriteria User Studies

Pu et al. [2011] conducted a user study to determine a set of recommendation quality criteria that accurately reflect the users' perception of a recommender system's usefulness. Of the beyond-accuracy objectives, diversity and novelty were included. Serendipity was discarded as it was considered too similar to novelty. The users were asked to find an information item using a recommender system of their choice and to answer a set of questions regarding the perceived qualities of the service. Having analyzed the correlations between answers, the authors validated a model consisting

of 32 criteria grouped into 15 categories. The results of the study showed the perceived usefulness of a recommender to be influenced by the perceived accuracy and novelty, and to a lesser extent by the perceived diversity.

Knijnenburg et al. [2012] proposed a framework for evaluating users' experience of recommender systems, including the perceived accuracy, satisfaction, choice difficulty, and diversity. The framework consists of a set of structurally related concepts including objective system aspects (e.g., the recommendation algorithm) and user characteristics (e.g., the user's age) that are connected to subjective user experiences (e.g., the perceived diversity). The authors proposed a set of questions to record the subjective user experiences and conducted a series of experiments to investigate the relationships between framework components. The results showed that diversification (implemented using the greedy reranking approach, see Algorithm 1) is perceived differently for different algorithms. For example, the users perceived recommendations of the k-NN algorithm with no diversification as more diverse than diversified recommendations of the same algorithm, while this was not observed for the factorization algorithm. When the users did perceive recommendations as diverse, this had a positive relationship with the perceived accuracy, the ease of choice, and consequently the overall satisfaction with the system.

Ekstrand et al. [2014] adapted the questions used by Knijnenburg et al. [2012] for a comparative study where users of the MovieLens recommender system were asked to compare pairs of movie recommendation lists and answer questions regarding the perceived accuracy, diversity, and novelty of the recommendation lists and their overall satisfaction. Three state-of-the-art algorithms were used for generating recommendations: an SVD factor model and both a user-user and an item-item collaborative filtering approach. To address the possible item familiarity effects, the authors limited the set of recommendable items to popular ones, thus avoiding recommendation lists with too many obscure items. The study results revealed that the users were equally satisfied with the SVD and item-item algorithms, while being less satisfied with the user-user algorithm. The perceived satisfaction with the recommendations was found to positively correlate with the perceived diversity and negatively with perceived novelty.

The observed negative influence of novelty on users' satisfaction seems to contradict the findings of Pu et al. [2011], who found novelty to positively influence the perceived system usefulness (and consequently users' satisfaction). The findings may differ due to different recommendation domains (Pu et al. conducted the survey using a number of online recommender services including Amazon, while Ekstrand et al. focused on movie recommendations). Another possible explanation lies in the different formulations of the novelty-related questions the users had to answer during the two studies. Pu et al. analyzed the perceived novelty by measuring the users' agreement with the statement "The recommender system helped me discover new products," while Ekstrand et al. compared the perceived novelty of two movie lists with such questions as "Which list has more movies you do not expect?" and "Which list has more movies you would not have thought to consider?" Therefore, in the first case, novelty feedback was gathered by means of a positive question, while in the second case, the negative tone of the survey questions may have tied negative user experiences to the measured objective, that is, novelty. This example shows the impact that the formulation of survey questions may have on the outcome of studies measuring users' perception of multiple recommendation qualities.

6.2. Beyond-Accuracy Objectives in Targeted User Studies

Targeted user studies are conducted to validate the usefulness of a certain technique, for example, a diversity-oriented algorithm, or a modification of the user interface optimized for diversity perception. Ziegler et al. [2005] evaluated their greedy

diversification technique on a BookCrossing dataset.[3] Given a list of book recommendations, each user was asked to evaluate the relevance of each recommendation, the diversity of the list of recommendations, and their overall satisfaction with the recommendations. The users were randomly assigned either a user-based or item-based CF recommender, and the diversification algorithm was based on comparing the genres of items (using a genre taxonomy-based metric). The results of the study showed that light diversification (changing up to four items in a list of 10 recommendations) positively influences user satisfaction with the item-based CF recommender. In the case of the user-based CF recommender, the results showed no measurable effect on satisfaction.

Celma [2009] evaluated the users' perception of item relevance and novelty in a music recommender. The author conducted a study with 288 Last.fm users who were asked to rate their familiarity and appreciation of the songs recommended by three algorithms—an item-based collaborative approach, a content-based approach, and a hybrid combination of the two. The results showed that the users perceived the recommendations of the content-based approach to be the most novel (i.e., they were least familiar with them), but also the least accurate (i.e., they assigned the lowest ratings to the tracks). Conversely, the collaborative approach was shown to produce the least novel but highest-rated recommendations. Celma hypothesizes that the low ratings of novel recommendations may be improved by providing explanations about why particular unknown songs are recommended.

Hu and Pu [2011] analyzed how a standard list-based user interface compares to a more organized interface that groups recommendations into categories in terms of perceived diversity. The authors conducted a within-subject user study with 20 participants in which each user viewed two versions of product recommendations ("customers who viewed this item also viewed")—one version showed the standard list of products; the other version showed groups of products in separate tabs (organized by brand or price range). The users provided feedback regarding the perceived categorical diversity (i.e., items being of different kinds) and item-to-item diversity (i.e., items being dissimilar to each other) as well as the perceived ease of use and usefulness of the system. While the results showed no significant difference in perceived item-to-item diversity, the perceived categorical diversity was shown to be larger in the second version of the interface, which also had a positive influence on the perceived ease of use and the usefulness of the system.

Willemsen et al. [2011] described a user study where diversification was based on latent item features in a matrix factorization model. A within-subjects study with 97 participants required each user to evaluate three lists of movie recommendations representing different levels of diversification: low-level, midlevel, and high-level diversity. For each of the three conditions, the perceived recommendation diversity and attractiveness were measured. The results showed that lists with high levels of diversity were perceived as most diverse by the users. Interestingly, the perceived attractiveness of recommendations increased from low- to midlevel diversification, but did not further increase for the high level of diversification. This result suggests that after a certain level of diversity is achieved, users may not appreciate further diversification.

Ge et al. analyzed the impact that the placement of items within a recommendation list has on users' perception of diversity. The authors conducted a pilot user study with 10 participants [Ge et al. 2011] and a later study with 52 participants [Ge et al. 2012]. The users were asked to evaluate the diversity of precomputed movie lists. The same lists were displayed to all the users. Each list contained movies from one genre, with a small number of different genre items inserted for diversity. The authors called the

[3]http://grouplens.org/datasets/book-crossing/.

items whose genre was different from the list's dominant genre "diverse items." Three experimental conditions were compared: inserting all "diverse items" at the end of the list, inserting them in the middle of the list, and distributing them throughout the list. The later study showed that distributing "diverse items" throughout the list or placing them together at the bottom of the list led to higher perceived diversity and higher surprise than placing them close to each other in the middle of the list. Moreover, having recognized the presence of "diverse items," users in the pilot study were interested in additional information about such items (possibly trying to understand why they were recommended). This result indicates the potential use of explanations in diversity-aware systems.

The importance of explanations in recommendation diversification is also mentioned by Castagnos et al. [2013], who conducted a user study with a specially created movie dataset consisting of around 500 movies, 3,000 users, and 173,000 ratings. The study involved 250 participants divided into five groups, with each group evaluating recommendations produced by one of the five algorithms—a baseline popularity approach, a content-based approach, an item-based collaborative filtering approach, and two collaborative filtering reranking techniques (variations of the greedy reranking technique, see Algorithm 1). The study results showed that, while the users positively perceived the diversification (movies suggested by the two diversity-aware techniques received the highest ratings on average), they had more confidence in recommendations produced by the least diverse approach—the content-based technique. Castagnos et al. explain this result in terms of the users' appreciation of recommendation transparency: the users were more confident when they clearly understood why a particular item was recommended (e.g., being very similar to a previously rated item). This result suggests that while diversity may be positively perceived by users, additional explanations may be important for improving the acceptance of such recommendations.

Zhang et al. [2012] performed a small-scale user study (21 participants) to evaluate a serendipity-enhancing recommender for Last.fm music artists (see also Section 3.2). Participants of the study were asked to provide six artists they like as a "seed" for the recommender and subsequently to evaluate two recommendation lists generated by a baseline recommender and the serendipity-enhancing version of the system. The perceived enjoyment (from "dislike the song" to "will definitely listen again") and serendipity (from "exactly what I listen to normally" to "something I would never have listened to otherwise") of the recommendations were measured on a 5-point Likert scale. The users' familiarity with the recommended artists was also recorded. The study results showed the users to perceive recommendations generated by the serendipity-enhancing system version as less enjoyable, but more serendipitous. The serendipity-enhancing version was also shown to provide more recommendations of novel artists (i.e., ones unknown to the users). Interestingly, despite providing less enjoyable recommendations, the serendipity-enhancing system version was preferred over the baseline system as the users were willing to sacrifice recommendation accuracy for the sake of discovering new interesting artists. An important question is whether users in domains where the cost of receiving inaccurate recommendations is higher (e.g., movie recommendations) would also be prepared to sacrifice accuracy for serendipity.

Finally, we observe that research discussed in this section deals with user studies in which participants are aware of their involvement in the experiments. While such studies may reveal important findings, they can only offer an approximation of the user behavior in real-life settings. There is a lack of reported A/B evaluation studies (i.e., online experiments where the users are unaware of their participation) that analyze the impact of beyond-accuracy objectives on user behavior.

7. OFFLINE ANALYSIS OF BEYOND-ACCURACY OBJECTIVES

Having reviewed research that addresses the definition, optimization, and measurement of the different beyond-accuracy objectives, we aim to further contribute to the beyond-accuracy recommendation research with a novel analysis of relationships between the different objectives. While existing research on diversity, novelty, serendipity, or coverage typically addresses one specific objective, we believe it is important to understand which objectives are corelated, which are in conflict, and how optimizing for one objective can affect the other objectives.

A number of previous works report on experiments where multiple beyond-accuracy objectives are measured in offline settings. Next we review their findings and position our work with respect to these previous efforts.

Ribeiro et al. [2012] proposed an approach for balancing recommendation accuracy, diversity, and novelty using a weighted combination of the predicted relevance scores that come from a number of different recommendation algorithms. The authors have measured the performance of various state-of-the-art recommendation approaches (including popularity-based, content-based, k-nearest-neighbor, and matrix factorization techniques) in terms of accuracy, diversity, and novelty. Offline evaluation results on MovieLens and Last.fm datasets showed that none of the algorithms dominates in all three objectives: on the MovieLens dataset, the SVD factor model with 50 factors provided the most accurate recommendations, the popularity-based approach the most diverse, and the SVD model with 150 factors the most novel. On Last.fm data, the factor model for implicit data generated the most accurate recommendations, the SVD model with 150 factors the most novel, and the user-based k-NN approach the most diverse. Combining predictions generated by the different algorithms resulted in hybrid solutions that performed similarly to the best algorithms in each individual objective (accuracy, diversity, or novelty), but better in the other two objectives.

Bellogín et al. [2013] evaluated the performance of a number of recommendation approaches that the authors grouped into three categories: rating-based techniques, content-based techniques (exploiting content labels), and social techniques (exploiting friendship relations between users). In addition to the traditional precision and recall metrics, the performance metrics included diversity (α-$nDCG$ metric [Clarke et al. 2008]), novelty, and coverage. Experiments on three datasets—Delicious, Last.fm, and MovieLens—showed the content-based approach to achieve the highest coverage and novelty on the Last.fm and Delicious datasets. Interestingly, on the MovieLens dataset, a user-based k-NN method provided the most novel recommendations. On Last.fm and Delicious data, the social recommenders were best in terms of diversity. (Social recommenders could not be applied to the MovieLens dataset as the data contains no user-to-user relations.) On the MovieLens dataset, content-based approaches were shown to outperform rating-based techniques in terms of α-$nDCG$ diversity.

Pampín et al. [2014] analyzed the performance of item-based and user-based k-NN approaches in terms of accuracy, diversity, and novelty. They conducted offline experiments on the MovieLens dataset with different values of the neighborhood size k for the user-based approach and a fixed value of $k = 300$ for the item-based approach. The evaluation results showed that at small values of k, the user-based approach provides more novel recommendations than the item-based approach, but the novelty decreases for larger values of k and matches the novelty of item-based recommendations at $k = 100$. The user-based approach was also shown to provide more diverse recommendations compared to the item-based approach, with diversity decreasing for larger values of k but remaining higher compared to the diversity of item-based recommendations (which were computed with $k = 300$).

Jannach et al. [2013] analyzed the tendency of various recommendation approaches to focus on certain parts of the item catalog. They evaluated recommendations in terms of coverage (measured as the aggregate number of items appearing in top-10 recommendation lists and as the Gini coefficient metric [Fleder and Hosanagar 2009]) and popularity bias (measured as the average rating value of recommended items and as the average number of ratings per item). The recommendation techniques that they evaluated included state-of-the-art neighbor-based and matrix factorization algorithms, a learning-to-rank approach for implicit rating data, and a content-based approach based on item labels. The evaluation results on the MovieLens dataset showed that the most accurate algorithm—the learning-to-rank approach—tends to focus on the popular items in the catalog, being also the worst in terms of novelty. Interestingly, the learning-to-rank approach performed well in terms of coverage, being second only to the content-based approach. The SVD matrix factorization approach also performed well in terms of coverage, while neighbor-based approaches achieved low coverage, showing a tendency to focus on a small portion of the item catalog. Overall, all rating-based algorithms were shown to suffer to some extent from popularity bias, the content-based approach being the only approach that was not biased toward popular items.

Compared to the previous efforts discussed, our work contains a broader set of performance metrics, covering the most popular recommender system beyond-accuracy objectives. For diversity, novelty, and coverage, we employed the metric definitions most widely encountered in the literature, while for the serendipity objective, we propose two alternative ways of measuring surprise, which constitutes the core component of serendipity.

Recently, Maksai et al. [2015] analyzed a wide range of beyond-accuracy metrics—multiple variants of diversity, novelty, coverage, and serendipity—in an offline setting. The authors suggested predicting the online performance of a news recommender (measured by the click-through rate) using the offline metric values. To identify which offline metrics are most likely to influence online performance of the system, Maksai et al. analyzed correlations between the metrics by measuring their values at equal time intervals on a news recommendation dataset. The obtained results indicate no strong correlations between the metrics, with the exception of a strong (positive) correlation between coverage (computed using Shannon entropy [Shani and Gunawardana 2011]) and serendipity (computed using the definition of Murakami et al. [2008]).

Our approach to measure metric correlations differs from that of Maksai et al.—rather than observing the change of metric values over time, we analyze how optimizing a recommender system for a specific objective affects other beyond-accuracy objectives. We address this problem by evaluating a number of greedy reranking approaches against the different beyond-accuracy metrics.

7.1. Reranking Approaches

In earlier sections of this article, we reviewed a number of the different approaches that have been proposed to enable recommender systems to generate not only accurate but also novel and surprising recommendations and diverse lists of recommendations. One approach relies on reranking the lists of recommendations that are generated using baseline algorithms [Smyth and McClave 2001; Ziegler et al. 2005; Kelly and Bridge 2006; Adomavicius and Kwon 2012]. Other approaches require the development of new recommendation models where accuracy and additional objectives are addressed simultaneously [Vargas et al. 2011; Oku and Hattori 2011; Hurley 2013; Su et al. 2013].

In our experiments, we chose to use the reranking approach, since it allows us to use existing state-of-the-art recommendation algorithms and allows us to explicitly control the tradeoff between recommendation accuracy and diversity, novelty, or serendipity. The reranking approach that we adopt follows the greedy algorithm described by

Smyth and McClave [2001] (see Section 2.2.1, Algorithm 1), where a list of candidate recommendations is reranked by greedily maximizing an objective function:

$$f_{obj}(i, R) = \alpha \cdot rel(i) + (1 - \alpha) \cdot obj(i, R). \tag{6}$$

The function combines recommendation accuracy and one of the beyond-accuracy objectives: given an item i, the item's predicted relevance, $rel(i)$, is combined with its diversity, novelty, or surprise score relative to the items already in the result list R, which we denote by $obj(i, R)$. To control the balance between accuracy and the alternative objectives, the $rel(i)$ and $obj(i, R)$ scores were standardized and the α parameter was set to 0.5 in all the experiments. In the following, we describe the different implementations of $obj(i, R)$ that were used in the experiments.

7.1.1. Diversity Reranking. We adopted the definition of diversity that is based on the average pairwise item distance, which is widely accepted in the RS literature [Smyth and McClave 2001; Ziegler et al. 2005; Kelly and Bridge 2006; Vargas and Castells 2011] (see Section 2.2.1, Equation (2)):

$$obj_{diversity}(i, R) = \frac{1}{|R|} \sum_{j \in R} \text{dist}(i, j). \tag{7}$$

As mentioned in Section 2.1, the item distance function $\text{dist}(i, j)$ has been defined using a variety of metrics. In our experiments, we evaluate two variants of the distance function—one based on item content labels and the other based on item ratings.

Content-Based Diversity. We employ the complement of the Jaccard similarity metric for comparing items described with a set of content labels (e.g., movies or artists labeled with genres):

$$\text{dist}(i, j) = 1 - \frac{|L_i \cap L_j|}{|L_i \cup L_j|}, \tag{8}$$

where L_i and L_j are the sets of labels describing items i and j, respectively.

Rating-Based Diversity. An alternative formulation of item distance is based on user-assigned ratings. We use the complement of the adjusted cosine similarity normalized to [0,1]:

$$\text{dist}(i, j) = \frac{1}{2} - \frac{\sum_{u \in U}(r_{ui} - \bar{r_i})(r_{uj} - \bar{r_j})}{2\sqrt{\sum_{u \in U}(r_{ui} - \bar{r_i})^2}\sqrt{\sum_{u \in U}(r_{uj} - \bar{r_j})^2}}, \tag{9}$$

where $\bar{r_i}$ and $\bar{r_j}$ are the average rating values for items i and j, respectively. Only users who rated both items are considered for the diversity computation.

7.1.2. Surprise Reranking. Our approach to measuring surprise is based on the intuition that a recommendation is surprising if it is unlike *any* item the user has seen before. We use the *lower-bound* item distance from the items in the user's profile as an indicator of surprise. We chose the lower-bound distance function rather than the more commonly used *average* distance since we believe that averaging the distance scores results in information loss, especially if the user has diverse items in his or her rating profile (see Kaminskas and Bridge [2014] for details).

We propose two alternative definitions of surprise, based on different item distance functions. The first metric exploits users' rating behavior to measure the likelihood for a pair of items to be seen by the same user. While this information is not direct evidence of item dissimilarity, it provides a reasonable approximation—items that are rarely observed together are likely to be different. The second metric employs a more straightforward item distance function, based on content labels.

Given the target item and the user's profile (a set of items rated by the user), both metrics produce a score that indicates the level of surprise the target item brings to the user. Note that, unlike some previous works [Vargas and Castells 2011; Adamopoulos and Tuzhilin 2014], we do not consider item relevance in our definitions of surprise. While relevance is an important component of serendipity, we leave it to be measured by dedicated accuracy metrics. Furthermore, neither metric is presently rank aware, although they could be adapted to discount items that appear lower in the recommendation list.

Co-occurrence-Based Surprise. The first definition is based on the probability for the item to be seen (i.e., rated) together with the items in the user's profile. To measure the pairwise co-occurrence of items, we employed normalized point-wise mutual information (PMI) [Bouma 2009], which measures the probability of observing specific outcomes of two independent random variables together. Given a pair of items i and j, we compute their PMI value as

$$\text{PMI}(i, j) = \log_2 \frac{p(i, j)}{p(i)p(j)} / - \log_2 p(i, j), \tag{10}$$

where $p(i)$ and $p(j)$ represent the probabilities for the items to be rated by any user, that is, $p(i) = \frac{|\{u \in U, r_{ui} \neq \emptyset\}|}{|U|}$, and $p(i, j)$ is the probability for the same user to rate both items, that is, $p(i, j) = \frac{|\{u \in U, r_{ui} \neq \emptyset \wedge r_{uj} \neq \emptyset\}|}{|U|}$. PMI values range from -1 (in the limit) to 1, with -1 meaning the two items are never rated together, 0 signifying independence of the items, and 1 meaning complete co-occurrence of the items.

In order to measure the surprise of a recommended item i, we compute its PMI with each item in the user's profile. Since higher values of PMI(i, j) signify higher co-occurrence of items i and j (and therefore low surprise of seeing the two items together), we take the complement of the PMI normalized to [0,1]. Taking the minimum of these values indicates the lower bound of the surprise perceived by the user when item i is recommended:

$$obj^{co-occ}_{surprise}(i) = \min_{j \in P} \frac{1 - \text{PMI}(i, j)}{2}, \tag{11}$$

where P is the user's profile (i.e., his or her set of rated items). (In this equation, and in Equations (12) and (13), we drop the parameter R, i.e., we write $obj^{co-occ}_{surprise}(i)$ rather than $obj^{co-occ}_{surprise}(i, R)$, since these metrics do not depend on the items already in the result list.)

Content-Based Surprise. Our second surprise metric is based on distance applied to item content labels:

$$obj^{cont}_{surprise}(i) = \min_{j \in P} \text{dist}(i, j), \tag{12}$$

where the distance is computed as the complement of Jaccard similarity (see Equation (8)).

Similarly to the co-occurrence-based definition, the distance is computed for all pairs consisting of the target item i and the items in the user's profile. Taking the minimum distance value as the overall surprise represents the lower bound of how surprising the item is with respect to the seen items.

7.1.3. Novelty Reranking. For novelty, we use the item's *self-information* or *inverse user frequency* [Zhou et al. 2010; Vargas and Castells 2011], which is the fraction of users

in the dataset who rated the item i:

$$obj_{novelty}(i) = -\log_2 \frac{|\{u \in U, r_{ui} \neq \emptyset\}|}{|U|}. \quad (13)$$

The logarithm is used to emphasize the novelty of the most rare items.

7.2. Experimental Setup

To study the relationships between the different beyond-accuracy objectives, we conducted a number of offline experiments using four state-of-the-art recommendation algorithms and the five variants of the greedy reranking approach described previously—two variants for both diversity (Equations (8) and (9)) and surprise (Equations (11) and (12)) and one for novelty (Equation (13)). In each experiment, a recommendation algorithm was used to generate a ranked list of candidate recommendations C ($|C| = 50$). Then, we reranked C using each of the five beyond-accuracy objectives (Equation (6), $\alpha = 0.5$). Finally, we obtained the list of top-N recommendations ($N = 10$ was used in all the experiments) from the reranked lists.

The value of $C = 50$ has been chosen to allow for a sufficiently large pool of candidate items while not slowing the performance significantly. Using the value of $\alpha = 0.5$ allowed a good balance between the predicted relevance and the beyond-accuracy objective (the scores of the two components in Equation (6) were standardized).

Next, we describe the evaluation methodology, performance metrics, datasets, and recommendation algorithms that were employed in the experiments.

7.2.1. Evaluation Methodology. In recent years, rating-based accuracy metrics for offline RS evaluations have been replaced by precision-oriented metrics that more closely reflect the users' interaction with the system—considering only a small set of top-ranked recommendations, ignoring the lower-ranked items [Bellogín et al. 2011]. In accordance with these state-of-the-art evaluation strategies, in this work we adopt the *one plus random* methodology [Koren 2008]. The methodology is based on randomly splitting each user's ratings to give a training set M and probe set P. The test set T is constructed by selecting all highly rated items (e.g., those having a five-star rating on a 1 to 5 scale) from the user's probe set P. Then, for each user u and for each test item (from T), predictions are computed for 1,000 random unrated items plus the one test item. The set of 1,001 items is ranked according to the recommender's predicted scores and the top-N recommendations are selected. If the test item is among the top-N items, we have a hit. The overall performance of the system—*recall*—is calculated as the ratio of the number of hits over the total number of test cases.

In this article, results were obtained using a slight modification of the methodology—rather than selecting *all* highly rated items for each user's test set T, we only used *one* randomly selected test item per user. This way equal importance is given to all test users, whereas the original methodology allows users with larger test profiles to have more impact on the evaluation results.[4]

The underlying assumption of the *one plus random* methodology that the 1,000 unseen items are irrelevant is clearly undervaluing the performance, as certain items among the 1,000 may be actually relevant for the user. However, we believe this methodology to be appropriate when measuring the beyond-accuracy objectives of recommendations as it involves items the user has not discovered (i.e., unrated items), whereas other offline evaluation strategies only employ items the user has had no trouble discovering (i.e., already-rated items).

[4]A subset of the experiments was also conducted using the full test sets of the users as per the methodology in Koren [2008], but no significant differences in the results were observed.

All experiment results reported in the following sections were computed using five-fold cross-validation with 80%/20% training/probe set split.

7.2.2. Performance Metrics. In addition to the *Recall* metric, for each test user's top-N recommendation list R, we compute the following beyond-accuracy metrics:

—Two variants of the diversity metric—the rating-based diversity and the content-based diversity:

$$Div_{ratings/cont}(R) = \frac{\sum_{i \in R} \sum_{j \in R \setminus \{i\}} \text{dist}(i, j)}{|R|(|R| - 1)}, \tag{14}$$

where for Div_{cont}, dist(i, j) is computed based on item content labels (Equation (8)), and for $Div_{ratings}$, dist(i, j) is computed using the rating-based item distance (Equation (9))

—Two variants of the surprise metric—the co-occurrence-based surprise and the content-based surprise:

$$S_{co-occ}(R) = \frac{1}{|R|} \sum_{i \in R} \min_{j \in P} \frac{1 - \text{PMI}(i, j)}{2} \tag{15}$$

$$S_{cont}(R) = \frac{1}{|R|} \sum_{i \in R} \min_{j \in P} \text{dist}(i, j), \tag{16}$$

where P is the target user's profile (i.e., the set of rated items), PMI(i, j) is computed using Equation (10), and dist(i, j) uses Equation (8)

—The *Novelty* metric computed as the average item self-information:

$$Novelty(R) = \frac{1}{nov_{max} \cdot |R|} \sum_{i \in R} -\log_2 \frac{|\{u \in U, r_{ui} \neq \emptyset\}|}{|U|}, \tag{17}$$

where U is the set of all users in the dataset and $nov_{max} = -\log_2 \frac{1}{|U|}$ is the maximal possible novelty value, which is used to normalize the novelty score of each individual item into [0,1].

The previous metric values are averaged across all test users.

Finally, we measure the *Coverage* metric as the aggregate number of distinct items appearing in top-N lists of all test users:

$$Coverage = |\cup_{u \in U} R_u|, \tag{18}$$

where R_u is the set of top-N recommendations generated for user u and U is the set of all test users.

Note that five of the performance metrics (the ones related to diversity, surprise, and novelty) correspond to the five reranking approaches described in Section 7.1. For convenience, we adapt the notation of each reranking approach to the corresponding metric. For instance, we refer to the rating-based diversity metric as $Div_{ratings}$ and we refer to the corresponding reranking approach (i.e., that reranks using this metric) as $Div^r_{ratings}$ (Section 7.1.1, Equations (7) and (9)).

The *Recall* and *Coverage* metrics do not have corresponding reranking approaches. Thus, in total, we have seven metrics and five reranking approaches.

7.2.3. Datasets. We tested the proposed beyond-accuracy reranking approaches on two benchmark datasets for offline recommender system evaluation—the MovieLens 1M dataset[5] and the Last.fm 1K dataset.[6]

The MovieLens dataset contains ∼1 million ratings, 6,040 users, and 3,706 movies. The movies are annotated using a vocabulary of 18 genres (on average 1.65 genres per movie). To obtain richer content descriptors for the movies, we additionally scraped IMDb plot keywords for each movie and kept those labels that appeared in the profiles of at least 10 movies. This resulted in an average of 60 labels per movie.

The Last.fm dataset contains the listening events for 992 users and more than 100K artists. As the dataset is extremely sparse, we cleaned the set of artists by leaving only those for which we could obtain at least three Last.fm tags (using the *artist.getTopTags* method of the Last.fm API[7]) and discarding artists who were listened to by fewer than 20 users. This resulted in 992 users and 7,280 artists, with a total of 500K ratings. The listening frequencies of the artists were transformed into ratings from 1 to 5 using the standard approach for converting frequency-based implicit feedback into numerical ratings [Celma 2009]. To avoid noisy data, we retrieved a maximum of the 10 most popular labels for every artist and kept the labels that appeared in the profiles of at least 10 artists. This resulted in eight labels per artist on average.

7.2.4. Recommendation Algorithms. The reranking approaches described in Section 7.1 were evaluated with four state-of-the-art recommendation algorithms: a pairwise learning-to-rank algorithm [Weston et al. 2010] (*LTR*), a *PureSVD* [Cremonesi et al. 2010] matrix factorization algorithm implemented using the *sparsesvd* library[8] (*MF*), and two *k*-nearest-neighbor algorithms—a user-based collaborative filtering method (*UB*) and an item-based collaborative filtering method (*IB*) [Desrosiers and Karypis 2011]. We note that to achieve an optimal accuracy on the Last.fm dataset (which consists of implicit feedback data converted to explicit numeric ratings), an algorithm designed for implicit feedback may be a better choice. However, since the goal of our experiments was not achieving the highest possible accuracy on both datasets but rather investigating the behavior of state-of-the-art algorithms with respect to beyond-accuracy metrics, we chose the four algorithms as representatives of the techniques most commonly discussed in the RS literature and evaluated the *same* set of algorithms with both datasets.

We optimized each algorithm's parameters (using a grid search strategy) to maximize recommendation accuracy (i.e., the *Recall* metric) as this is the standard practice in RS research. For the *LTR* algorithm, we optimized the regularization constant $C \in \{1, 10, 100, 1000\}$, the learning rate $\gamma \in \{0.01, 0.001, 0.0001\}$, and the number of factors $f \in \{25, 50, 75, 100\}$; for the *MF* algorithm, we optimized the number of factors $f \in \{25, 50, 75, 100, 150, 175, 200, 250\}$; and for the k-*NN* algorithms, the neighborhood size $k \in [20, 250]$. The selected parameter values for each algorithm/dataset combination are as follows:

—*LTR*: $f = 25$, $C = 1{,}000$, $\gamma = 0.01$ for the MovieLens dataset, and $f = 50$, $C = 10$, $\gamma = 0.01$ for the Last.fm dataset
—*MF*: $f = 25$ for both datasets
—*UB*: $k = 150$ for the MovieLens dataset, and $k = 20$ for the Last.fm dataset
—*IB*: $k = 60$ for both datasets

[5]http://grouplens.org/datasets/movielens/.
[6]http://www.dtic.upf.edu/∼ocelma/MusicRecommendationDataset/lastfm-1K.html.
[7]http://www.last.fm/api/show/artist.getTopTags.
[8]https://pypi.python.org/pypi/sparsesvd/.

For comparison, we note that Jannach et al. [2013] in their analysis of beyond-accuracy objectives used item-based and user-based k-NN algorithms with $k = 100$ and an SVD matrix factorization algorithm with the number of factors $f = 50$. Bellogín et al. [2013] conducted experiments with $k = 15$ for both k-NN algorithms and $f = 50$ for the matrix factorization algorithm. Pampín et al. [2014] found the accuracy of the user-based k-NN algorithm to improve with increasing k until $k = 90$ and to thereafter remain constant at least until $k = 200$. (They did not optimize the item-based algorithm and fixed the value at $k = 300$ for their experiments.)

In the following sections, we report our findings and compare the results with the previous works discussed in Section 7. It is important to note that any reported differences should not be treated as definite conclusions but rather as indications for further research, as they may be influenced by a number of factors, such as the differences in preprocessing of the datasets, tuning of the algorithm parameters, and evaluation methodologies [Said and Bellogín 2014]. In particular, the *one plus random* methodology adopted in our experiments has been shown to favor approaches recommending more popular items [Cremonesi et al. 2010; Jannach et al. 2015b]. Consequently, optimizing a recommendation algorithm's parameters using this methodology may result in the algorithm's configuration being more popularity oriented compared to the parameter settings used in other studies. However, we believe that the possible popularity biases of individual algorithms do not invalidate the findings of this research since we focus on the relative comparison of reranking strategies applied to the output of individual algorithms.

7.3. Results and Discussion

We conducted two main sets of experiments. One was aimed at comparing the performance of the four recommender algorithms (optimized for recall, as discussed in the previous section) in terms of the different beyond-accuracy performance metrics (Section 7.3.1). The other was aimed at evaluating the five reranking approaches: we used each reranking of the results of each of the four recommendation algorithms and recorded all performance measures (Section 7.3.2). Furthermore, we report initial observations regarding the influence of algorithm parameters on the performance metrics and reranking effectiveness (Section 7.3.3).

7.3.1. Comparison of Recommendation Algorithms. Figures 1 and 2 show the results obtained for each of the four recommendation algorithms on the MovieLens and Last.fm datasets, respectively.

On the MovieLens dataset (Figure 1), the *LTR* and *MF* algorithms show a similar performance: both are the best in accuracy (*Recall*) and diversity (Div_{cont} and $Div_{ratings}$) and lose to the k-NN algorithms in surprise (S_{cont} and S_{co-occ}). However, the *LTR* algorithm significantly outperforms the *MF* algorithm in *Novelty* and *Coverage*, which indicates a tendency of the matrix factorization algorithm to focus on popular items (i.e., ones with low novelty).

We compare these results with the findings of Jannach et al. [2013], who used the MovieLens dataset (although rather than using the 1 million rating set, Jannach et al. used a subset of the 10 million rating set) and a different learning-to-rank approach (the Bayesian Personalized Ranking algorithm, which was designed for implicit feedback data). Jannach et al. report that both learning-to-rank and matrix factorization achieve high catalog coverage but perform poorly in terms of novelty (the learning-to-rank algorithm being particularly vulnerable to the popularity bias). In our results, we observe a more direct link between high coverage and novelty—the *LTR* algorithm doing well at both and the *MF* algorithm showing inferior performance.

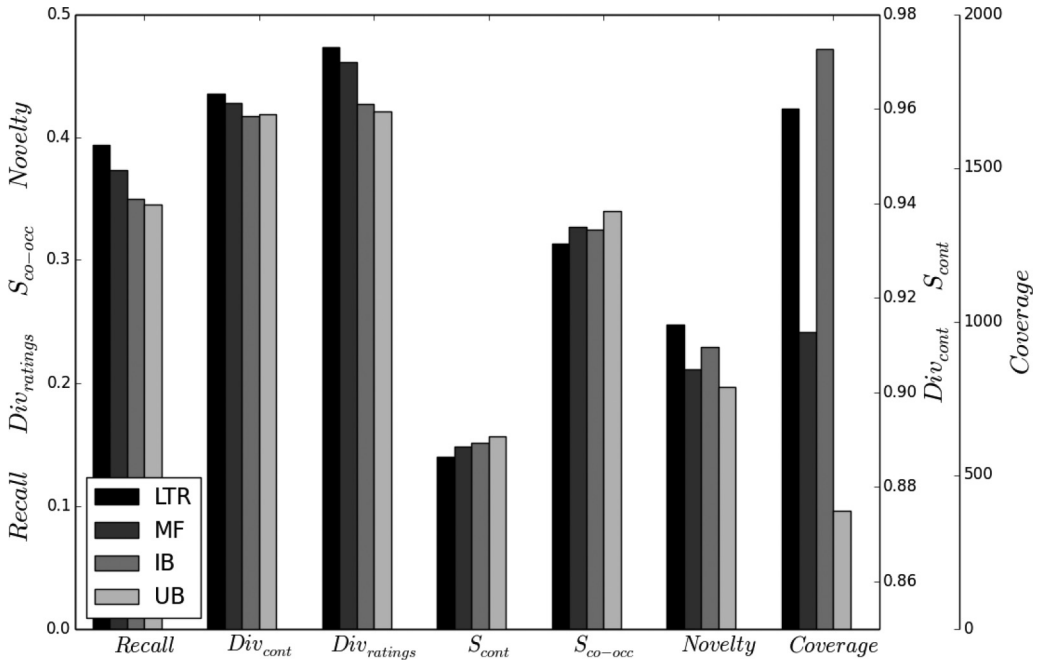

Fig. 1. Metric comparison for the evaluated algorithms (MovieLens dataset).

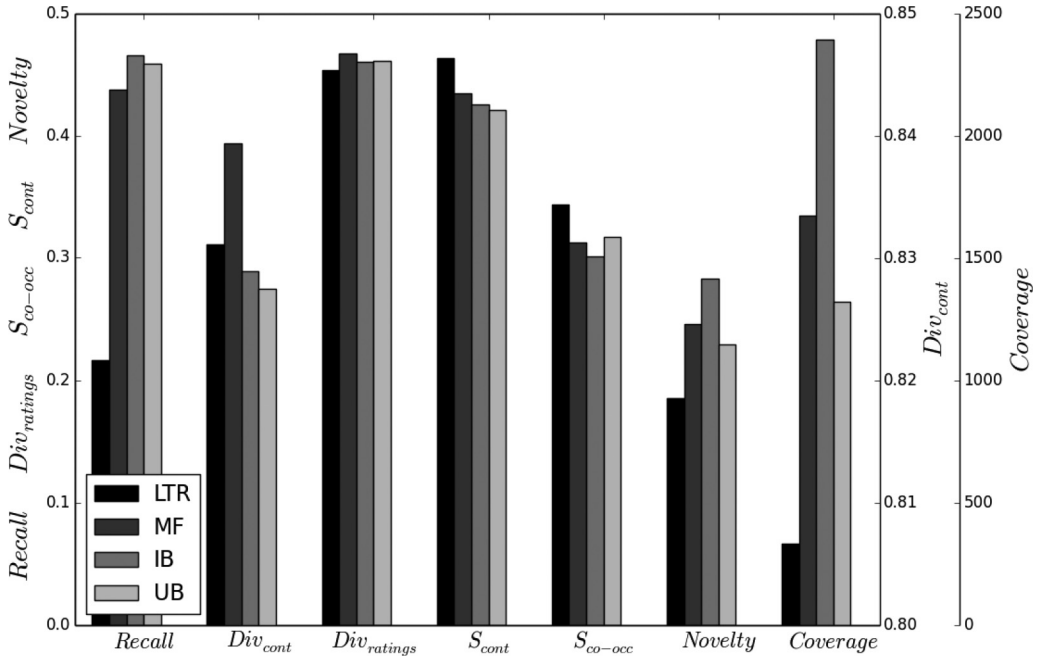

Fig. 2. Metric comparison for the evaluated algorithms (Last.fm dataset).

The *IB* and *UB* k-NN algorithms show a similar performance in terms of *Recall*, Div_{cont}, and $Div_{ratings}$ metrics, both losing to the *LTR* and *MF* algorithms. The *IB* k-NN algorithm achieves second-best performance in *Novelty* (losing only to the *LTR* algorithm) and is the best in *Coverage*. On the other hand, the *UB* algorithm is generating the most surprising recommendations (S_{cont} and S_{co-occ}) but shows the worst performance among the four algorithms in terms of *Novelty* and *Coverage*, which indicates a tendency to recommend popular items.

The observed performance of the *UB* algorithm is in contrast with results reported by Bellogín et al. [2013], who found the user-based k-NN algorithm to generate the most novel recommendations on the MovieLens dataset (compared to an item-based k-NN and an SVD algorithm). This difference may be explained by their neighborhood size ($k = 15$), as smaller neighborhood size corresponds to higher novelty (see Section 7.3.3). Jannach et al. found the *UB* algorithm to perform poorly in terms of coverage, which matches our findings. They also report a tendency of the *UB* algorithm to recommend either the most popular or the most novel items (with approximately equal frequency)—a result that could not be confirmed by our experiments, since we measure the algorithm's novelty by averaging the novelty scores of items in recommendation lists (Equation (17)).

On the Last.fm dataset (Figure 2), the results show a few differences compared to the MovieLens dataset. The main difference in the results is the behavior of the *LTR* algorithm, which now loses to other algorithms in terms of *Recall*, $Div_{ratings}$, *Novelty*, and *Coverage*. Moreover, the *LTR* algorithm shows the best results in surprise (S_{cont} and S_{co-occ}).

The inferior performance of the *LTR* algorithm in terms of *Recall* could be caused by the recommender algorithms we chose to use and the rating data in the dataset—implicit feedback converted to numeric ratings (Section 7.2.4).

The *MF* algorithm is the best in diversity (particularly Div_{cont}) and is also second best in *Novelty* and *Coverage*, losing to the *IB* algorithm. Differently from the MovieLens results, the *MF* algorithm also performs better than the k-NN algorithms in terms of content-based surprise S_{cont}.

Interestingly, the results on Last.fm data also show both k-NN algorithms to achieve the best *Recall*, slightly outperforming the *MF* algorithm. This is in line with previous works showing that, when evaluating top-N recommendations (particularly using the *one plus random* methodology), the accuracy of simple techniques may be similar to that of the more advanced algorithms [Cremonesi et al. 2010; Jannach et al. 2013].

We compare our results on Last.fm data to the findings of Bellogín et al. [2013], who evaluated (among other techniques) item-based and user-based k-NN algorithms as well as an SVD factorization algorithm on a Last.fm dataset (the authors built a dedicated dataset with 1.9K users and 17.6K artists). Results reported by Bellogín et al. confirm the high novelty and coverage achieved by the item-based k-NN algorithm. However, the authors also report the user-based k-NN algorithm to achieve second-best novelty and the best coverage results. As said earlier, this result can be explained by their small neighborhood size ($k = 15$).

7.3.2. Comparison of Reranking Approaches. Figure 3 shows the performance measure values obtained using the different reranking approaches with the *MF* algorithm on the MovieLens dataset. Analogous sets of results were obtained for every dataset-algorithm combination (eight sets of results in total). To better describe the results, in this section, we outline the findings that are largely consistent across both datasets. A full set of result figures is presented in the appendix.

In Figure 3 (and all the figures in the appendix), each individual chart (a) through (g) shows the values for one performance metric (named on the y-axis), with the different

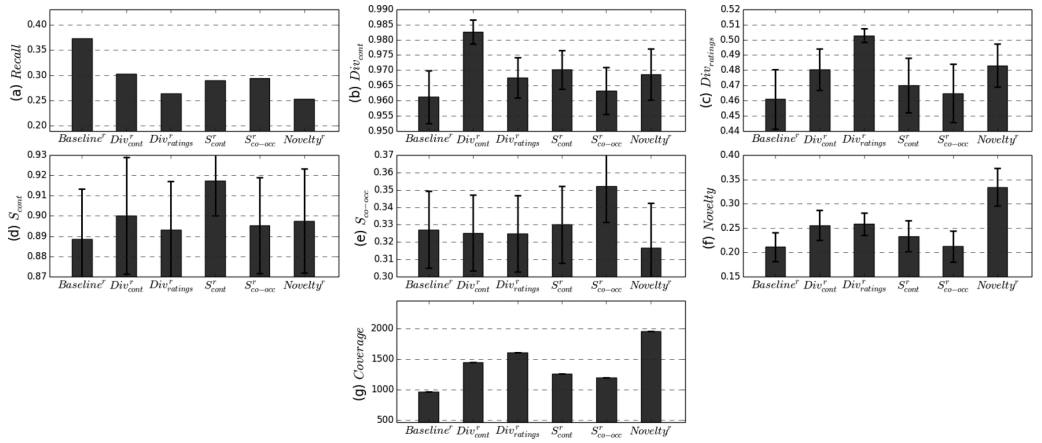

Fig. 3. Metric values for the different reranking approaches with the *MF* algorithm (MovieLens data).

reranking approaches displayed along the x-axis. The $Baseline^r$ approach corresponds to the original recommendation ranking generated by the respective algorithm. In other words, in the baseline there is no reranking.

As said in Section 7.2.2, each performance metric except for *Recall* and *Coverage* corresponds to a reranking approach. For instance, the diversity reranking $Div^r_{ratings}$ is the approach that reranks the list of candidate recommendations according to the rating-based diversity objective (see Section 7.1.1) and therefore corresponds to the $Div_{ratings}$ metric (see Equation (14)).

As expected, *Recall* has its highest value when using the $Baseline^r$ ranking for each algorithm and is lowered using any of the reranking approaches (chart (a) in each figure). This illustrates the well-known tradeoff between recommendation accuracy and beyond-accuracy objectives. For the diversity, surprise, and novelty metrics (charts (b)–(f) in each figure), the highest values are achieved using the corresponding reranking approaches, which is the expected outcome. As said in Section 7, our main focus in these experiments was observing how each beyond-accuracy metric is affected by reranking approaches that are not directly optimizing the metric. These observations allow us to identify positive or negative correlations between the different beyond-accuracy objectives: if a reranking approach significantly improves a metric compared to the baseline ranking, we can assume there exists a positive correlation between the reranking objective and the metric; on the other hand, if a reranking approach results in a value for the metric that is lower than the baseline, we assume a negative correlation between the reranking objective and the metric.

Here we outline the discovered correlations between the different objectives. As said earlier, the results for the Last.fm dataset are largely consistent with those obtained with the MovieLens data (the few notable exceptions are mentioned later).

—Reranking the recommendations for novelty (i.e., using the $Novelty^r$ reranking) hurts accuracy the most (see Figure 3, chart (a)). This is not surprising in the offline evaluation setting, as any offline evaluation methodology is (to a certain extent) biased toward popular items: user ratings in the test set are more likely to belong to popular items. It is worth noting that in three cases (*MF*, *UB*, and *IB* algorithms on Last.fm dataset), $Novelty^r$ reranking has the second-worst *Recall* performance, with the largest accuracy loss shown for the S^r_{cont} reranking approach (see the appendix, Figure 12(a)).

—Rating-based diversity is positively correlated with novelty, since reranking $Novelty^r$ positively influences the rating-based diversity metric $Div_{ratings}$ (Figure 3(c)) and vice versa—reranking $Div^r_{ratings}$ positively influences the *Novelty* metric (Figure 3(f)).
—There is a positive correlation between the content-based diversity and the content-based surprise (Figures 3(b) and 3(d)). This correlation might be explained by the fact that both Div_{cont} and S_{cont} metrics as well as the corresponding reranking approaches Div^r_{cont} and S^r_{cont} use the Jaccard item distance function based on the content labels (Equation (8)).
—A negative correlation is observed among the co-occurrence-based surprise and novelty, since reranking $Novelty^r$ results in S_{co-occ} values lower than the baseline (Figure 3(e)) and reranking S^r_{co-occ} results in *Novelty* values slightly lower than the baseline (appendix, Figure 8(f)). This indicates that the S_{co-occ} metric is scoring the long-tail items lower than the popular items. The finding confirms previous results: the metric is sensitive to item popularity. This is because its core component—pointwise mutual information (Equation (10))—is sensitive to pairs of rare items (see Kaminskas and Bridge [2014] for more discussion). For future use, the metric may need to be modified to avoid such bias.
—*Coverage* is positively influenced by the $Div^r_{ratings}$ and $Novelty^r$ rerankings (Figure 3(g)). The positive relationship between coverage and novelty is expected, as discussed in Section 5, while the positive influence of rating-based diversity may be linked to its correlation with novelty (see earlier). Interestingly, for the *UB* and *IB* k-NN algorithms (appendix, Figure 8(g)), the highest *Coverage* value is achieved when reranking for rating-based diversity $Div^r_{ratings}$ (with $Novelty^r$ reranking achieving the second-best value).
—An exception to the previous finding is the negative influence on *Coverage* obtained by the $Novelty^r$ reranking with the *LTR* algorithm on the Last.fm dataset (appendix, Figure 10(g)). This may indicate that the *LTR* algorithm tends to focus on a particular section of the long-tail item distribution in the Last.fm dataset; however, more detailed analysis of the issue is needed before reaching definite conclusions.

As discussed in Section 7.2.4, the evaluation methodology adopted in our experiments may favor popularity-oriented algorithm configurations. This may have an impact on the results obtained by the reranking strategies, as certain algorithms can have more popular items among the reranking candidates. However, we are only interested in comparison of reranking techniques within the same algorithm recommendations. To conduct a cross-algorithm comparison of reranking strategies, additional popularity bias metrics (such as the average rating of recommended items [Jannach et al. 2015b]) could be employed. We leave this to future work.

7.3.3. The Impact of Algorithm Parameters. We also conducted experiments aimed at investigating the influence of recommendation algorithm parameters on the different performance metrics. For the *UB* and *IB* k-NN algorithms, we computed all the performance metrics while varying the value of the neighborhood size parameter k. For the *MF* algorithm, we varied the number of factors f. We did not include the learning-to-rank (*LTR*) algorithm in this set of experiments, since the algorithm's three parameters result in a much larger parameter search space.

Due to the large number of experiment runs required for the different parameter values, the results reported in this section were computed with a random sample of 1,000 users on the MovieLens dataset (rather than the full set of 6,040 users).

Figures 4 and 5 show results for the different parameter values of the *MF* and *UB* algorithms, respectively.

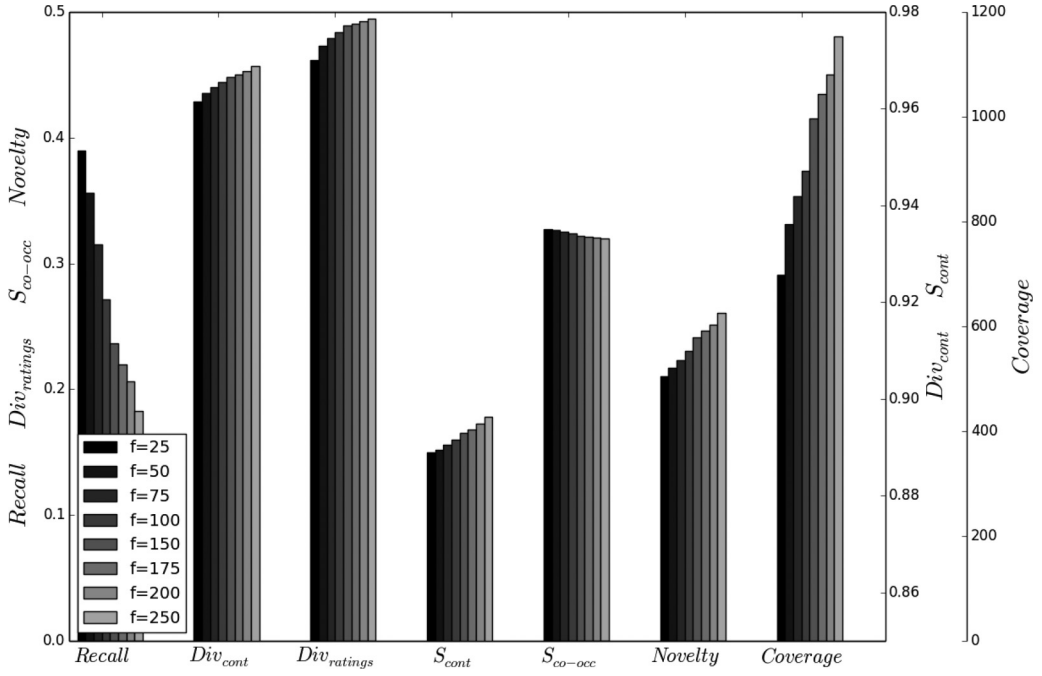

Fig. 4. The influence of the number of factors f on performance of the MF algorithm (MovieLens dataset).

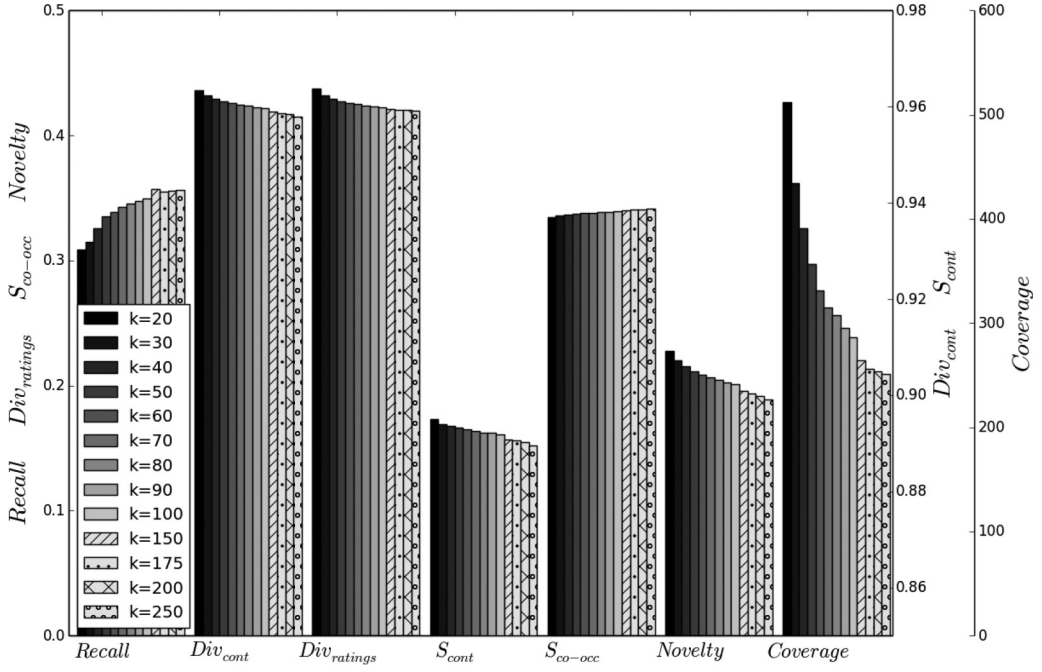

Fig. 5. The influence of the neighborhood size k on performance of the user-based k-NN algorithm (MovieLens dataset).

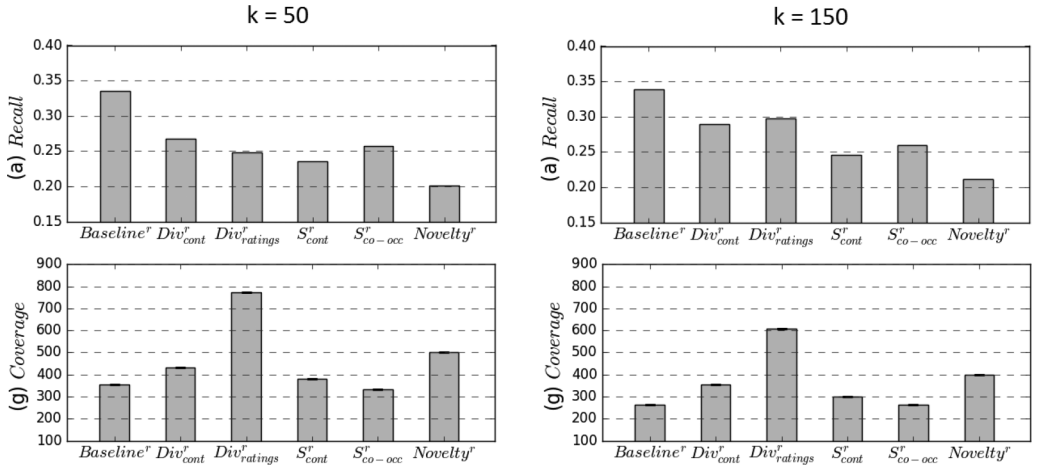

Fig. 6. Recall and coverage values obtained with the reranking approaches on two versions the *UB* algorithm: $k = 50$ and $k = 150$ (MovieLens data).

—For the *MF* algorithm, the results show that increasing the number of factors leads to a loss in accuracy (*Recall*). However, the beyond-accuracy performance metrics increase for higher f values. For instance, *Coverage* goes up from approximately 700 items for $f = 25$ to almost $1,200$ items for $f = 250$. The increase is also evident for the diversity metrics, *Novelty*, and content-based surprise S_{cont}. The co-occurrence-based surprise S_{co-occ} decreases slightly for higher f values, which may be linked to the metric's sensitivity to item popularity.
—For the *UB* algorithm, the results present an "inverted" picture: higher k values significantly decrease *Coverage* as well as the diversity metrics, *Novelty*, and S_{cont}, while S_{co-occ} values are slightly increasing. The *Recall* is increasing with the value of k, with the maximum reached at $k = 150$. Further increase of k does not improve the performance.
—For *IB*, the trend of decreasing *Novelty* and *Coverage* values for larger k values could be observed. However, we do not show this in a figure, as the observed impact of k on the metrics was much less pronounced. For instance, *Coverage* decreased from approximately 900 items for $k = 20$ to approximately 800 items for $k = 250$.

We note that the decreasing coverage values for larger neighborhood sizes in both k-NN approaches may seem counterintuitive as a larger user/item neighborhood leads to more items being considered for recommendation. However, this trend can only be measured using the "prediction coverage" metric (i.e., the ratio of items for which prediction can be made, see Section 5.1). Since we focus on the coverage of items that appear in the top-N recommendation lists, we obtain results that are caused by more popular items appearing among the top recommendations for larger neighborhood sizes.

An aspect of beyond-accuracy optimization that we did not fully address in our experiments is the influence of recommendation algorithm parameters on the effectiveness of reranking approaches. We observed interesting changes in the *Coverage* results for the different values of neighborhood size k with the *UB* algorithm (on MovieLens data). When using the best-performing (in terms of *Coverage* value) reranking $Div^r_{ratings}$ with the neighborhood size $k = 50$, we achieved a 30% higher coverage compared to that of $k = 150$ (Figure 6, bottom charts). While the difference in *Recall* for the $Baseline^r$ rankings of *UB* with $k = 50$ and $k = 150$ is approximately 0.01 (see Figure 6, upper

charts), it increases to approximately 0.05 for the $Div^{r}_{ratings}$ rankings. Although more analysis is needed to investigate this tradeoff, it is likely that a 0.05 loss in accuracy is a price worth paying for a 30% increase in recommendation coverage.

A detailed analysis of the impact of algorithm parameters on beyond-accuracy objectives is out of the scope of this article. We refer interested readers to Jannach et al. [2015b], where a number of state-of-the-art algorithms (with different parameter configurations) are analyzed with respect to various recommendation metrics (including popularity and concentration bias).

8. DISCUSSION AND CONCLUSIONS

In this article, we have reviewed the state-of-the-art research on beyond-accuracy objectives in recommender systems. We have focused on the four most widely discussed objectives—diversity, serendipity, novelty, and coverage. For each objective, we reviewed the relevant definitions found in the literature and methods for optimizing each objective.

Furthermore, we conducted offline experiments aimed at evaluating how the state-of-the-art recommendation algorithms perform in terms of the beyond-accuracy objectives and at studying the relationships between the objectives themselves. We have implemented a number of optimization strategies for improving diversity, serendipity, and novelty and investigated how optimizing each objective affects recommendation accuracy and beyond-accuracy metrics.

The main goal of this work was to provide a reference point for further research into improving the different beyond-accuracy qualities of recommender systems. We aimed both to survey the existing literature and to identify important relationships between the different objectives.

There are still many interesting challenges to address in this research area. We believe the following research directions to be of particular importance:

Evaluation of Beyond-Accuracy Objectives. As stated in Section 6, offline evaluation is limited when it comes to understanding the real impact of beyond-accuracy objectives on the users' experience. The gap between offline metrics and the users' perception of recommendation qualities has been exemplified in a recent study [Said et al. 2013], which showed that two algorithms recommending an almost disjoint set of items and obtaining significantly different accuracy scores in offline settings were perceived as equally useful by participants of a user study.

Even when performing user studies, a number of factors such as the domain of recommended items, the type of survey questions, or item familiarity effects can influence the results. For instance, it has been shown that item familiarity has a strong correlation with user appreciation of recommendations [Jannach et al. 2015a]. Ultimately, no results will be complete without conducting A/B experiments, where the users would be unaware of their involvement in the evaluation.

Adaptivity of Beyond-Accuracy Objectives. Another important challenge in beyond-accuracy research is developing optimization solutions that are adapted to specific recommendation domains, since different items may require different levels of recommendation diversity or novelty. For instance, the same novelty-enhancing algorithm may not suit both a movie recommender (where obvious recommendations are not desired) and a music streaming service (where well-known items may be among the desirable recommendations) [Kapoor et al. 2015].

An equally important challenge is to tailor the beyond-accuracy optimization to the needs or preferences of individual users. Solutions that adapt to users' needs or preferences for diversity, serendipity, or novelty may do so implicitly, for example, based

on the users' rating behavior [Shi et al. 2012] or the type of consumed items [Oh et al. 2011; Kapoor et al. 2015]. Alternatively, users may be given explicit control over their recommendations, for example, choosing to see more (or fewer) popular items [Harper et al. 2015].

Additionally, the solutions may be improved by ensuring transparency of beyond-accuracy recommendations. For instance, it has been argued that user acceptance of diversification may suffer if no explanation of the diversity level is provided [Castagnos et al. 2013].

APPENDIX

The appendix contains the figures of results discussed in Section 7.3.2. Figures 7 through 9 show the performance of the reranking approaches for the *LTR*, *IB*, and *UB* algorithms on the MovieLens dataset.

Figures 10 through 13 show the results for the *LTR*, *MF*, *IB*, and *UB* algorithms on the Last.fm dataset.

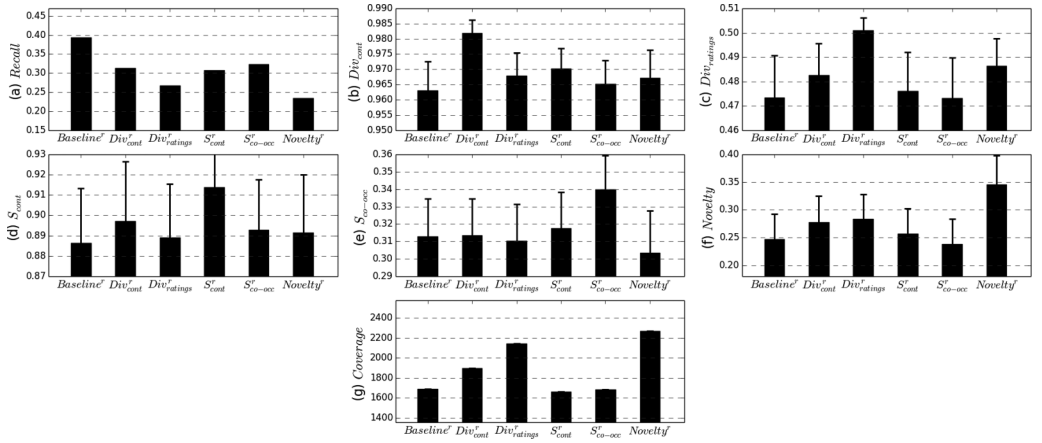

Fig. 7. Metric values for the different reranking approaches with the *LTR* algorithm (Movielens data).

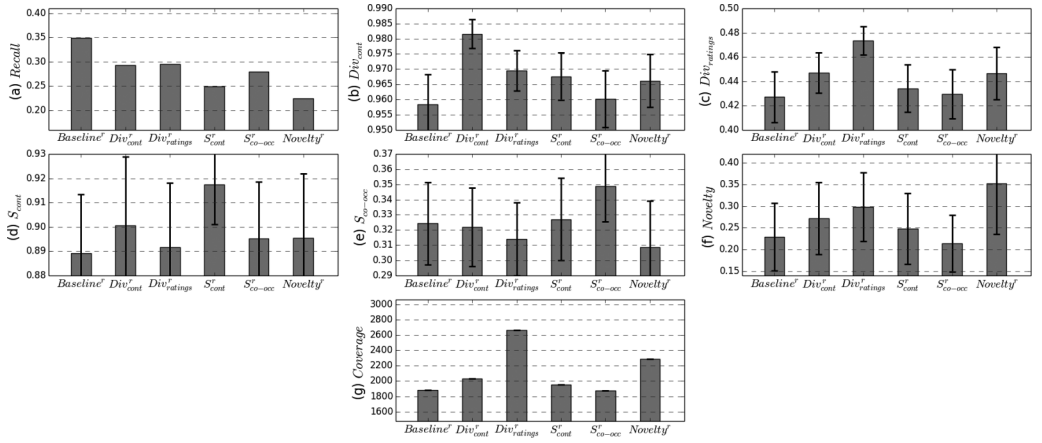

Fig. 8. Metric values for the different reranking approaches with the *IB* algorithm (MovieLens data).

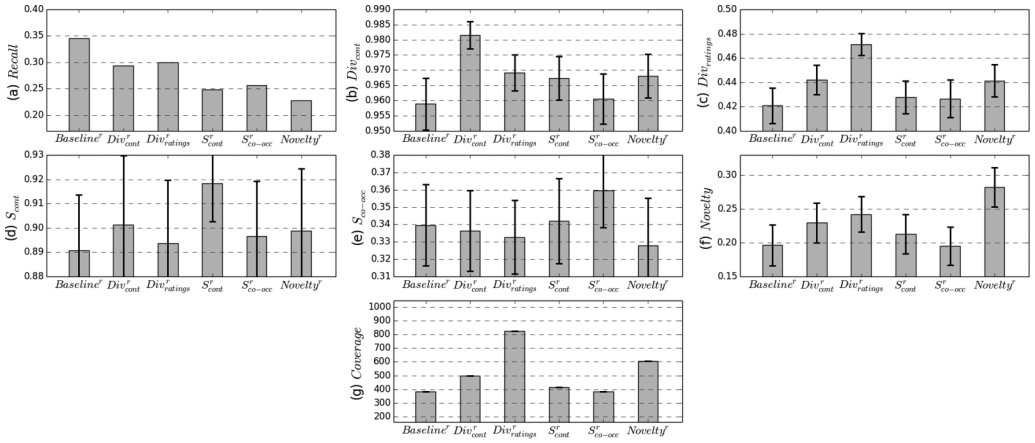

Fig. 9. Metric values for the different reranking approaches with the *UB* algorithm (Movielens data).

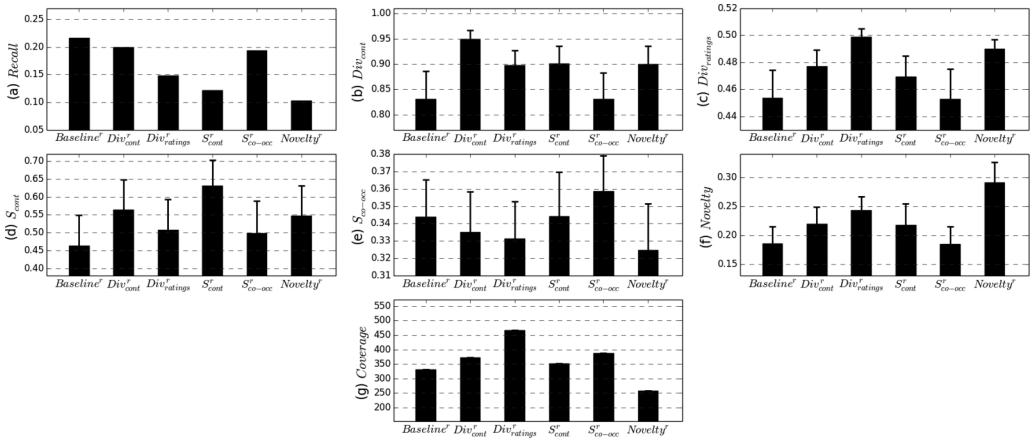

Fig. 10. Metric values against the different reranking approaches with the *LTR* algorithm (Last.fm data).

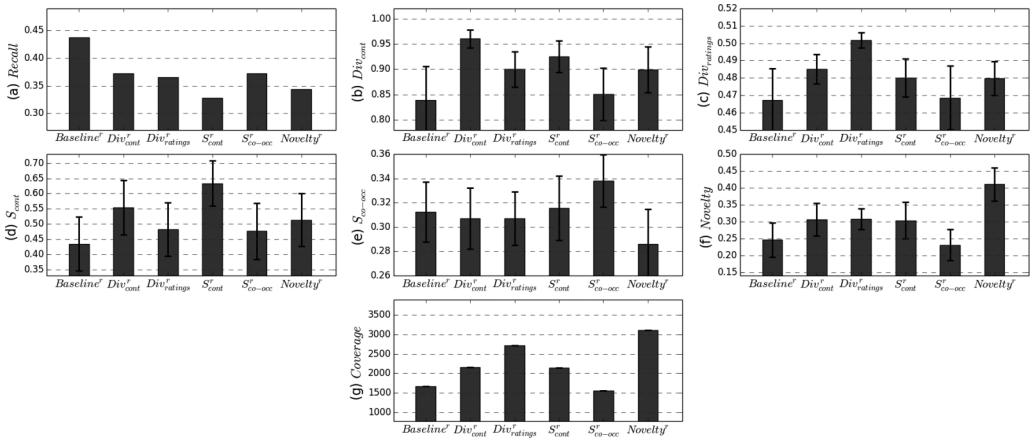

Fig. 11. Metric values against the different reranking approaches with the *MF* algorithm (Last.fm data).

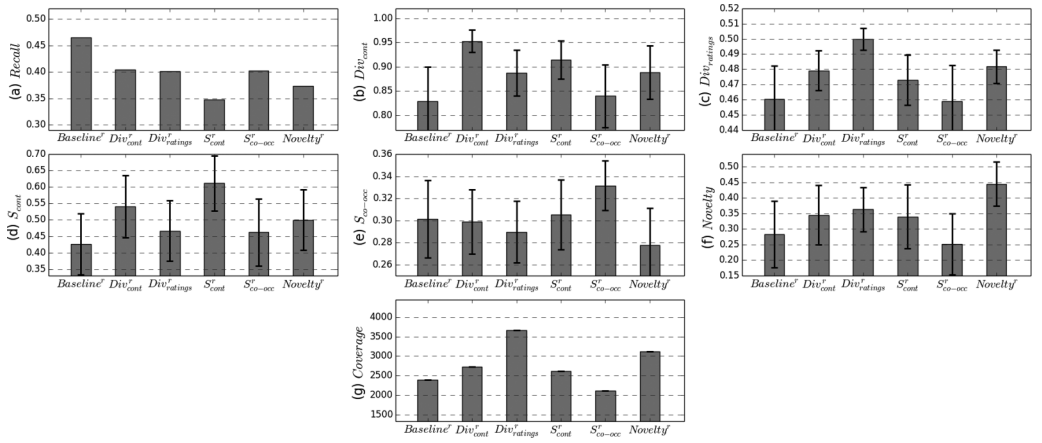

Fig. 12. Metric values against the different reranking approaches with the *IB* algorithm (Last.fm data).

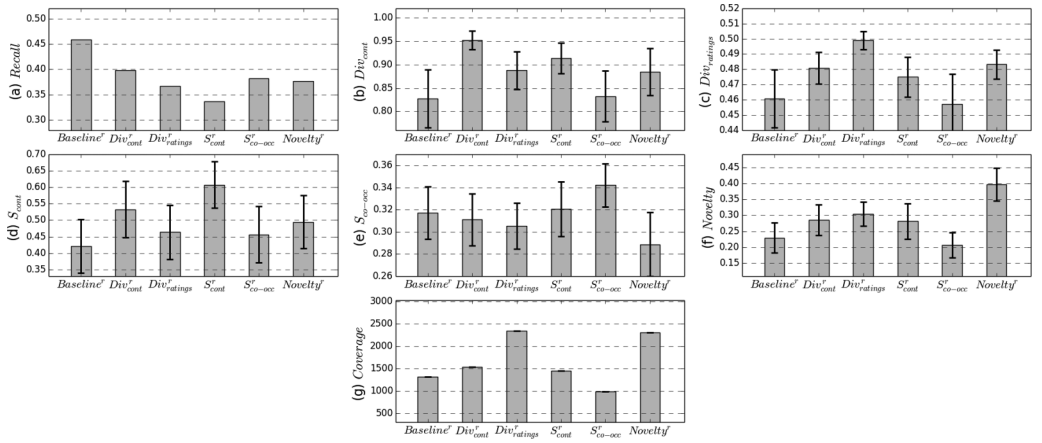

Fig. 13. Metric values against the different reranking approaches with the *UB* algorithm (Last.fm data).

REFERENCES

Panagiotis Adamopoulos and Alexander Tuzhilin. 2014. On unexpectedness in recommender systems: Or how to better expect the unexpected. *ACM Transactions on Intelligent Systems and Technology (TIST)* 5, 4 (2014), 54.

Gediminas Adomavicius and YoungOk Kwon. 2011. Maximizing aggregate recommendation diversity: A graph-theoretic approach. In *Proceedings of the 1st International Workshop on Novelty and Diversity in Recommender Systems (DiveRS'11)*. 3–10.

Gediminas Adomavicius and YoungOk Kwon. 2012. Improving aggregate recommendation diversity using ranking-based techniques. *IEEE Transactions on Knowledge and Data Engineering* 24, 5 (2012), 896–911.

Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. 2009. Diversifying search results. In *Proceedings of the 2nd ACM International Conference on Web Search and Data Mining*. 5–14.

Chris Anderson. 2006. *The Long Tail: Why the Future of Business Is Selling Less of More*.

Paul André, M. C. Schraefel, Jaime Teevan, and Susan T. Dumais. 2009. Discovery is never by chance: Designing for (un)serendipity. In *Proceedings of the 7th ACM Conference on Creativity and Cognition*. 305–314.

Ricardo Baeza-Yates and Berthier Ribeiro-Neto. 1999. *Modern Information Retrieval*. ACM Press / Addison-Wesley.

Andrea Barraza-Urbina, Benjamin Heitmann, Conor Hayes, and Angela Carrillo-Ramos. 2015. XPLODIV: An exploitation-exploration aware diversification approach for recommender systems. In *Proceedings of the 28th International Florida Artificial Intelligence Research Society Conference*.

Alejandro Bellogín, Iván Cantador, and Pablo Castells. 2013. A comparative study of heterogeneous item recommendations in social systems. *Information Sciences* 221 (2013), 142–169.

Alejandro Bellogín, Iván Cantador, Fernando Díez, Pablo Castells, and Enrique Chavarriaga. 2013. An empirical comparison of social, collaborative filtering, and hybrid recommenders. *ACM Transactions on Intelligent Systems and Technology (TIST)* 4, 1 (2013), 14.

Alejandro Bellogín, Pablo Castells, and Ivan Cantador. 2011. Precision-oriented evaluation of recommender systems: An algorithmic comparison. In *Proceedings of the 5th ACM Conference on Recommender Systems*. 333–336.

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. In *Proceedings of the Conference of the German Society for Computational Linguistics and Language Technology*. 31–40.

Fidel Cacheda, Víctor Carneiro, Diego Fernández, and Vreixo Formoso. 2011. Comparison of collaborative filtering algorithms: Limitations of current techniques and proposals for scalable, high-performance recommender systems. *ACM Transactions on the Web (TWEB)* 5, 1 (2011), 2.

José Campos and Antonio Dias de Figueiredo. 2001. Searching the unsearchable: Inducing serendipitous insights. In *Proceedings of the Workshop Program at the 4th International Conference on Case-Based Reasoning*. 159–164.

Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 335–336.

Sylvain Castagnos, Armelle Brun, and Anne Boyer. 2013. When diversity is needed... but not expected! In *Proceedings of the 3rd International Conference on Advances in Information Mining and Management*. 44–50.

Pablo Castells, Neil J. Hurley, and Saul Vargas. 2015. Novelty and diversity in recommender systems. In *Recommender Systems Handbook*. 881–918.

Òscar Celma. 2009. *Music Recommendation and Discovery in the Long Tail*. Ph.D. Dissertation. Universitat Pompeu Fabra.

Charles L. A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. 2008. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 659–666.

Paolo Cremonesi, Yehuda Koren, and Roberto Turrin. 2010. Performance of recommender algorithms on top-n recommendation tasks. In *Proceedings of the 4th ACM Conference on Recommender Systems*. 39–46.

Van Dang and W. Bruce Croft. 2012. Diversity by proportionality: An election-based approach to search result diversification. In *Proceedings of the 35th international ACM SIGIR Conference on Research and Development in Information Retrieval*. 65–74.

Christian Desrosiers and George Karypis. 2011. A comprehensive survey of neighborhood-based recommendation methods. In *Recommender Systems Handbook*, F. Ricci et al. (Eds.). 107–144.

Michael D. Ekstrand, F. Maxwell Harper, Martijn C. Willemsen, and Joseph A. Konstan. 2014. User perception of differences in recommender algorithms. In *Proceedings of the 8th ACM Conference on Recommender Systems*. 161–168.

Michael D. Ekstrand, John T. Riedl, and Joseph A. Konstan. 2011. Collaborative filtering recommender systems. *Foundations and Trends in Human-Computer Interaction* 4, 2 (2011), 81–173.

Daniel Fleder and Kartik Hosanagar. 2009. Blockbuster culture's next rise or fall: The impact of recommender systems on sales diversity. *Management Science* 55, 5 (2009), 697–712.

Allen Foster and Nigel Ford. 2003. Serendipity and information seeking: An empirical study. *Journal of Documentation* 59, 3 (2003), 321–340.

Meadhbh Foster and Mark Keane. 2013. Surprise: You've got some explaining to do. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society*. 2321–2326.

Mouzhi Ge, Carla Delgado-Battenfeld, and Dietmar Jannach. 2010. Beyond accuracy: Evaluating recommender systems by coverage and serendipity. In *Proceedings of the 4th ACM Conference on Recommender Systems*. 257–260.

Mouzhi Ge, Fatih Gedikli, and Dietmar Jannach. 2011. Placing high-diversity items in top-n recommendation lists. In *Proceedings of the Workshop on Intelligent Techniques for Web Personalization and Recommender Systems (ITWP'11)*. 65–68.

Mouzhi Ge, Dietmar Jannach, and Fatih Gedikli. 2012. Bringing diversity to recommendation lists–an analysis of the placement of diverse items. In *International Conference on Enterprise Information Systems*. 293–305.

Asela Gunawardana and Guy Shani. 2009. A survey of accuracy evaluation metrics of recommendation tasks. *Journal of Machine Learning Research* 10 (2009), 2935–2962.

F. Maxwell Harper, Funing Xu, Harmanpreet Kaur, Kyle Condiff, Shuo Chang, and Loren Terveen. 2015. Putting users in control of their recommendations. In *Proceedings of the 9th ACM Conference on Recommender Systems*. 3–10.

Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl. 2004. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems* 22, 1 (2004), 5–53.

Yoshinori Hijikata, Takuya Shimizu, and Shogo Nishida. 2009. Discovery-oriented collaborative filtering for improving user satisfaction. In *Proceedings of the 14th International Conference on Intelligent User Interfaces*. 67–76.

Rong Hu and Pearl Pu. 2011. Enhancing recommendation diversity with organization interfaces. In *Proceedings of the 16th International Conference on Intelligent User Interfaces*. 347–350.

Neil J. Hurley. 2013. Personalised ranking with diversity. In *Proceedings of the 7th ACM Conference on Recommender Systems*. 379–382.

Leo Iaquinta, Marco de Gemmis, Pasquale Lops, Giovanni Semeraro, Michele Filannino, and Piero Molino. 2008. Introducing serendipity in a content-based recommender system. In *Proceedings of the 8th International Conference on Hybrid Intelligent Systems*. 168–173.

Masayuki Ishikawa, Peter Geczy, Noriaki Izumi, and Takahira Yamaguchi. 2008. Long tail recommender utilizing information diffusion theory. In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. Vol 1. 785–788.

Tamas Jambor and Jun Wang. 2010. Optimizing multiple objectives in collaborative filtering. In *Proceedings of the 4th ACM Conference on Recommender Systems*. 55–62.

Dietmar Jannach, Lukas Lerche, Fatih Gedikli, and Geoffray Bonnin. 2013. What recommenders recommend–an analysis of accuracy, popularity, and sales diversity effects. In *User Modeling, Adaptation, and Personalization*. 25–37.

Dietmar Jannach, Lukas Lerche, and Michael Jugovac. 2015a. Item familiarity effects in user-centric evaluations of recommender systems. In *Poster Proceedings of the 9th ACM Conference on Recommender Systems*.

Dietmar Jannach, Lukas Lerche, Iman Kamehkhosh, and Michael Jugovac. 2015b. What recommenders recommend: An analysis of recommendation biases and possible countermeasures. *User Modeling and User-Adapted Interaction* 25, 5 (2015), 427–491.

Marius Kaminskas and Derek Bridge. 2014. Measuring surprise in recommender systems. In *Proceedings of the ACM RecSys Workshop on Recommender Systems Evaluation: Dimensions and Design (REDD'14)*.

Komal Kapoor, Vikas Kumar, Loren Terveen, Joseph A. Konstan, and Paul Schrater. 2015. I like to explore sometimes: Adapting to dynamic user novelty preferences. In *Proceedings of the 9th ACM Conference on Recommender Systems*. 19–26.

John Paul Kelly and Derek Bridge. 2006. Enhancing the diversity of conversational collaborative recommendations: A comparison. *Artificial Intelligence Review* 25, 1–2 (2006), 79–95.

Bart P. Knijnenburg, Martijn C. Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. 2012. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction* 22, 4–5 (2012), 441–504.

Yehuda Koren. 2008. Factorization meets the neighborhood: A multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 426–434.

Neal Lathia, Stephen Hailes, Licia Capra, and Xavier Amatriain. 2010. Temporal diversity in recommender systems. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 210–217.

Jian-Guo Liu, Kerui Shi, and Qiang Guo. 2012. Solving the accuracy-diversity dilemma via directed random walks. *Physical Review E* 85, 1 (2012), 016118.

Andrii Maksai, Florent Garcin, and Boi Faltings. 2015. Predicting online performance of news recommender systems through Richer evaluation metrics. In *Proceedings of the 9th ACM Conference on Recommender Systems*. 179–186.

Harry Markowitz. 1952. Portfolio selection. *Journal of Finance* 7, 1 (1952), 77–91.

Abigail McBirnie. 2008. Seeking serendipity: The paradox of control. In *Aslib Proceedings*. Vol. 60. 600–618.

Lorraine McGinty and Barry Smyth. 2003. On the role of diversity in conversational recommender systems. In *Proceedings of the 5th International Conference on Case-Based Reasoning*. 276–290.

Sean M. McNee, John Riedl, and Joseph A. Konstan. 2006. Being accurate is not enough: How accuracy metrics have hurt recommender systems. In *CHI'06 Extended Abstracts on Human Factors in Computing Systems*. 1097–1101.

Wulf-Uwe Meyer, Rainer Reisenzein, and Achim Schützwohl. 1997. Toward a process analysis of emotions: The case of surprise. *Motivation and Emotion* 21, 3 (1997), 251–274.

Tomoko Murakami, Koichiro Mori, and Ryohei Orihara. 2008. Metrics for evaluating the serendipity of recommendation lists. In *Proceedings of the 2007 Conference on New Frontiers in Artificial Intelligence*. 40–46.

Makoto Nakatsuji, Yasuhiro Fujiwara, Akimichi Tanaka, Toshio Uchiyama, Ko Fujimura, and Toru Ishida. 2010. Classical music for rock fans? Novel recommendations for expanding user interests. In *Proceedings of the 19th ACM Conference on Information and Knowledge Management*. 949–958.

Jinoh Oh, Sun Park, Hwanjo Yu, Min Song, and Seung-Taek Park. 2011. Novel recommendation based on personal popularity tendency. In *Proceedings of the 2011 IEEE 11th International Conference on Data Mining (ICDM'11)*. 507–516.

Kenta Oku and Fumio Hattori. 2011. Fusion-based recommender system for improving serendipity. In *Proceedings of the Workshop on Novelty and Diversity in Recommender Systems (DiveRS'11)*. 19–25.

Kensuke Onuma, Hanghang Tong, and Christos Faloutsos. 2009. TANGENT: A novel, "surprise me," recommendation algorithm. In *Proceedings of the 15th ACM Conference on Knowledge Discovery and Data Mining*. 657–666.

Humberto Jesús Corona Pampín, Houssem Jerbi, and Michael P. OMahony. 2014. Evaluating the relative performance of neighbourhood-based recommender systems. In *Proceedings of the 3rd Spanish Conference on Information Retrieval*.

Yoon-Joo Park and Alexander Tuzhilin. 2008. The long tail of recommender systems and how to leverage it. In *Proceedings of the 2nd ACM Conference on Recommender Systems*. 11–18.

Pearl Pu, Li Chen, and Rong Hu. 2011. A user-centric evaluation framework for recommender systems. In *Proceedings of the 5th ACM Conference on Recommender Systems*. 157–164.

Marco Tulio Ribeiro, Anisio Lacerda, Adriano Veloso, and Nivio Ziviani. 2012. Pareto-efficient hybridization for multi-objective recommender systems. In *Proceedings of the 6th ACM Conference on Recommender Systems*. 19–26.

Alan Said and Alejandro Bellogín. 2014. Comparative recommender system evaluation: Benchmarking recommendation frameworks. In *Proceedings of the 8th ACM Conference on Recommender Systems*. 129–136.

Alan Said, Ben Fields, Brijnesh J. Jain, and Sahin Albayrak. 2013. User-centric evaluation of a k-furthest neighbor collaborative filtering recommender algorithm. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*. 1399–1408.

Guy Shani and Asela Gunawardana. 2011. Evaluating recommendation systems. In *Recommender Systems Handbook*, F. Ricci et al. (Eds.). 257–297.

Lei Shi. 2013. Trading-off among accuracy, similarity, diversity, and long-tail: A graph-based recommendation approach. In *Proceedings of the 7th ACM Conference on Recommender Systems*. 57–64.

Yue Shi, Martha Larson, and Alan Hanjalic. 2010. List-wise learning to rank with matrix factorization for collaborative filtering. In *Proceedings of the 4th ACM Conference on Recommender Systems*. 269–272.

Yue Shi, Martha Larson, and Alan Hanjalic. 2014. Collaborative filtering beyond the user-item matrix: A survey of the state of the art and future challenges. *ACM Computing Surveys (CSUR)* 47, 1 (2014), 3.

Yue Shi, Xiaoxue Zhao, Jun Wang, Martha Larson, and Alan Hanjalic. 2012. Adaptive diversification of recommendation results via latent factor portfolio. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 175–184.

Barry Smyth and Paul McClave. 2001. Similarity vs. diversity. In *Proceedings of the 4th International Conference on Case-Based Reasoning*. 347–361.

Ruilong Su, Li'Ang Yin, Kailong Chen, and Yong Yu. 2013. Set-oriented personalized ranking for diversified top-n recommendation. In *Proceedings of the 7th ACM Conference on Recommender Systems*. 415–418.

Kirsten Swearingen and Rashmi Sinha. 2001. Beyond algorithms: An HCI perspective on recommender systems. In *ACM SIGIR 2001 Workshop on Recommender Systems*. 393–408.

Elaine G. Toms. 2000. Serendipitous information retrieval. In *Proceedings of 1st DELOS Workshop on Information Seeking, Searching and Querying in Digital Libraries*. 11–14.

Pek Van Andel. 1994. Anatomy of the unsought finding. Serendipity: Origin, history, domains, traditions, appearances, patterns and programmability. *British Journal for the Philosophy of Science* 45, 2 (1994), 631–648.

Saúl Vargas, Linas Baltrunas, Alexandros Karatzoglou, and Pablo Castells. 2014. Coverage, redundancy and size-awareness in genre diversity for recommender systems. In *Proceedings of the 8th ACM Conference on Recommender Systems*. 209–216.

Saúl Vargas and Pablo Castells. 2011. Rank and relevance in novelty and diversity metrics for recommender systems. In *Proceedings of the 5th ACM Conference on Recommender Systems*. 109–116.

Saúl Vargas and Pablo Castells. 2014. Improving sales diversity by recommending users to items. In *Proceedings of the 8th ACM Conference on Recommender Systems*. 145–152.

Saul Vargas, Pablo Castells, and David Vallet. 2011. Intent-oriented diversity in recommender systems. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1211–1212.

Jun Wang and Jianhan Zhu. 2009. Portfolio theory of information retrieval. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 115–122.

Jason Weston, Samy Bengio, and Nicolas Usunier. 2010. Large scale image annotation: Learning to rank with joint word-image embeddings. *Machine Learning* 81, 1 (2010), 21–35.

Martijn C. Willemsen, Bart P. Knijnenburg, Mark P. Graus, Linda C. M. Velter-Bremmers, and Kai Fu. 2011. Using latent features diversification to reduce choice difficulty in recommendation lists. *RecSys* 11 (2011), 14–20.

Yan Yang and Jian Z. Li. 2005. Interest-based recommendation in digital library. *Journal of Computer Science* 1, 1 (2005), 40.

Cong Yu, Laks Lakshmanan, and Sihem Amer-Yahia. 2009. Recommendation diversification using explanations. In *Proceedings of the 25th IEEE International Conference on Data Engineering (ICDE'09)*. 1299–1302.

Liang Zhang. 2013. The definition of novelty in recommendation system. *Journal of Engineering Science and Technology Review* 6, 3 (2013), 141–145.

Mi Zhang and Neil Hurley. 2008. Avoiding monotony: Improving the diversity of recommendation lists. In *Proceedings of the 2nd ACM Conference on Recommender Systems*. 123–130.

Yi Zhang, Jamie Callan, and Thomas Minka. 2002. Novelty and redundancy detection in adaptive filtering. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 81–88.

Yuan Cao Zhang, Diarmuid Ó Séaghdha, Daniele Quercia, and Tamas Jambor. 2012. Auralist: Introducing serendipity into music recommendation. In *Proceedings of the 5th ACM Conference on Web Search and Data Mining*. 13–22.

Tao Zhou, Zoltán Kuscsik, Jian-Guo Liu, Matúš Medo, Joseph Rushton Wakeling, and Yi-Cheng Zhang. 2010. Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences* 107, 10 (2010), 4511–4515.

Cai-Nicolas Ziegler, Sean M. McNee, Joseph A. Konstan, and Georg Lausen. 2005. Improving recommendation lists through topic diversification. In *Proceedings of the 14th International Conference on the World Wide Web*. 22–32.

Received November 2015; revised August 2016; accepted September 2016