International Journal of Machine Learning and Cybernetics (2019) 10:813–831
https://doi.org/10.1007/s13042-017-0762-9

ORIGINAL ARTICLE

# How good your recommender system is? A survey on evaluations in recommendation

**Thiago Silveira**[1] · **Min Zhang**[1] · **Xiao Lin**[1] · **Yiqun Liu**[1] · **Shaoping Ma**[1]

Received: 16 May 2017 / Accepted: 5 December 2017 / Published online: 14 December 2017


**Abstract**
Recommender Systems have become a very useful tool for a large variety of domains. Researchers have been attempting to improve their algorithms in order to issue better predictions to the users. However, one of the current challenges in the area refers to how to properly evaluate the predictions generated by a recommender system. In the extent of offline evaluations, some traditional concepts of evaluation have been explored, such as accuracy, Root Mean Square Error and P@N for top-k recommendations. In recent years, more research have proposed some new concepts such as novelty, diversity and serendipity. These concepts have been addressed with the goal to satisfy the users' requirements. Numerous definitions and metrics have been proposed in previous work. On the absence of a specific summarization on evaluations of recommendation combining traditional metrics and recent progresses, this paper surveys and organizes the main research that present definitions about concepts and propose metrics or strategies to evaluate recommendations. In addition, this survey also settles the relationship between the concepts, categorizes them according to their objectives and suggests potential future topics on user satisfaction.

**Keywords** Recommender system · Evaluation · Novelty · Diversity · Serendipity · Unexpectedness

## 1 Introduction

Recommender Systems (RSs) have been largely studied for the past decade and have shown to be suitable for many scenarios. On the arrival of the internet and the era of e-commerces, companies are opting for having a RS as an attempt to boost sales. RSs provide predictions of items that the user may find interesting to purchase [2], in which most algorithms for this purpose focus on providing recommendations that fit the preferences of the user.

RSs have shown to be useful for users and business. Users suffer from what is called the paradox of choice. Having many options to choose from lead to more difficulty in effectively making a choice [31]. Since e-commerces have a vast amount of items, users face a complication in finding what they desire. Therefore, RSs can help users, since good predictions can reduce the search space for the user, facilitating their decision making process [29]. Moreover, it is also advantageous for business, because it represents enlargement of sales. Specifically, RSs are able to increase sales of niche items. Popular items are visible to the users anyway, however niche items would not be very likely visible by all users, but personalized recommendations can find the right items for the right users [2, 6, 31].

There are plenty of examples of companies that use RSs. For instance, Amazon and many e-commerces have adopted the use of recommendation engines. Other services such as Netflix, Youtube and LastFm also use recommender systems.

In this way, researchers have been trying to design better recommenders that can suit the tastes of the users, stimulating them to purchase more. Personalized and non-personalized recommenders have been proposed [29]. In non-personalized recommenders, no users' information is used to make predictions. Though, personalized recommenders require users' past consumption information in order to issue recommendations. Personalized recommendations are more likely to suit users needs as they are based on the users' data [20].

In the search for a suitable recommendation algorithm, a question raises: *how good a recommender system is?* A methodology of evaluation is necessary in order to compare RSs. Historically, evaluations have been performed in online

✉ Min Zhang
zhuangzq16@mails.tsinghua.edu.cn; z-m@tsinghua.edu.cn

[1] Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science and Technology, Tsinghua University, Beijing, China

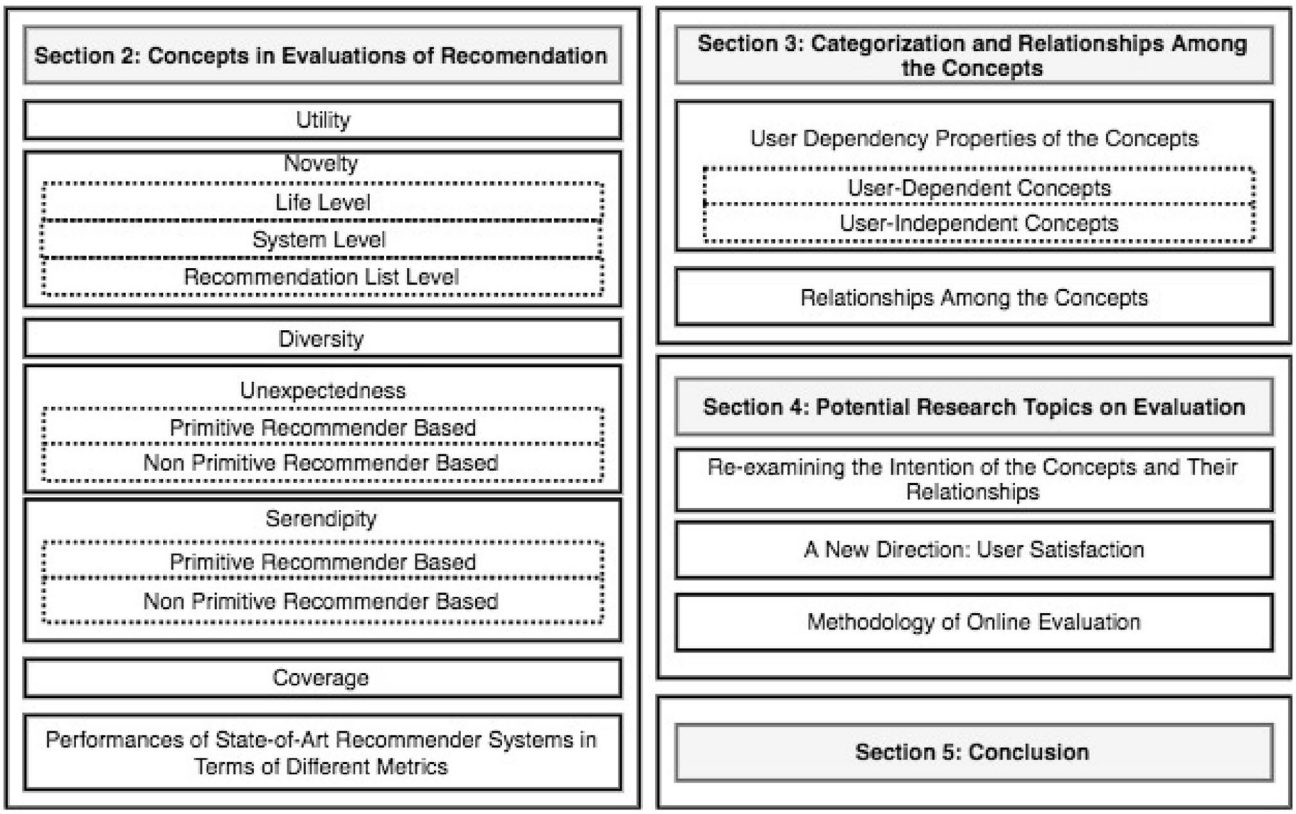


**Fig. 1** Overview and brief description of the survey's structure

and offline experiments. Online experiments involve issuing recommendations and then querying the users about how they rate the items [29]. Offline experiments do not require real users, instead part of the data is used to train the algorithm, while another sample is used to test the predictions regarding the users tastes [29].

Online evaluation is most desired, since it can provide accurate results of how good our system is with real users [29]. However, experiments with users are usually costly, then, many researchers opt for offline evaluations, instead. Early work in the area used to evaluate recommendations with machine learning metrics. For instance, accuracy-based metrics are widely adopted. These metrics capture the utility of the predictions, however they usually ignore users' desires for novelty and variety [1, 14]. Therefore, researchers have been not only evaluating usefulness of RSs, but also considering other concepts. Recent work have been worried about concepts such as novelty, diversity and surprise.

Evaluation of RSs is not trivial and authors have been proposing definitions and metrics for concepts of evaluation. Therefore, this paper's main goal is to survey evaluations of recommender systems. Section 2 inspects distinct concepts of evaluations of RSs, surveying work that researched about six concepts: *utility, novelty, diversity, unexpectedness, serendipity* and *coverage*. In Sect. 3, we present a categorization of the concepts and how they relate to each other. Finally, Sect. 4 discuss about possible future researches on assessment of recommendations. The Fig. 1 shows a roadmap illustrating the sections' contents.

## 2 Concepts in evaluations of recommendation

In the context of recommendation, researchers and professionals in RSs are concerned with user satisfaction, so that the predictions can provide the more value to the user. The reason is that RSs must be useful to the user, not only suggesting them to consume "more of the same". Researchers worry about users' interaction and consumption experience in the system. Recently, researchers have been attempting to solve this problem by evaluating different concepts of evaluation, rather than simply use predictive accuracy and machine learning techniques [14]. The performance of the suggestions provided by a RSs should be measured by the value it can generate to the user [14]. There are many concepts regarding evaluation of recommendations, such as coverage, novelty, diversity and surprise of recommendations have been evaluated by different researches.

**Table 1** Symbols used in the metrics

| Symbol | Meaning |
|---|---|
| $R_u$ | List of recommendations for user $u$ |
| $U$ | Set of users |
| $I$ | Set of items |
| $H_u$ | History of consumption of the user |
| $F_i$ | Set of features for item $i$ |
| $PM_u$ | Primitive recommender |
| $E_u$ | Expected items |
| $USEFUL_u$ | Useful items for user $u$ |
| $C_u$ | Total number of items the user consumed $R_u$ |
| $util(R_u)$ | Utility notation |
| $nov(R_u)$ | Novelty notation |
| $div(R_u)$ | Diversity notation |
| $unexp(R_u)$ | Unexpectedness notation |
| $ser(R_u)$ | Serendipity notation |
| $cov$ | Coverage notation |

This methodology of evaluation in recommendation has different names in the literature. Kotkov et al. [20, 21] use the word *concept* referring to novelty, relevance, serendipity. However, distinct names have been used. For instance, Adamopoulos et al. [1] use the word *dimensions* to refer to improvements that can increase the performance and usefulness of the recommendation list for users. Additionally, Herlocker et al. [14] prefer the name *measures of recommender system evaluation*, referring to coverage, confidence, novelty and learning rate. Besides having different names, we state the word *concept* to refer to different aspects for assessment of RSs.

In this section, we examine previous work in the literature that research about distinct concepts for evaluating recommendation in RS. Most existing concepts can be summarized into six different ones: *utility, novelty, diversity, unexpectedness, serendipity* and *coverage*. We investigate the main definitions and metrics for each one of them. For the sake of organization, the symbols used in all the metrics are standardized and summarized in Table 1. In addition, the Table 3 in the appendix contains a summarization of the metrics, categorization and their advantages and issues.

There are other concepts besides the six ones aforementioned, however, due to space limitation and lack of research about them, we decided to focus on the most studied ones. Nevertheless, some concepts that are not mentioned in this paper are: trust, risk, robustness, privacy, adaptability and scalability [29]. Even though the handbook developed by Ricci et al. [29] contains a review about the concepts of evaluation, our survey is focused on the evaluation concepts and metrics; moreover, this survey includes recent updates on metrics for unexpectedness and serendipity which were published after the review developed by Ricci et al. [29].

## 2.1 Utility

Utility has been mentioned in the literature in many names, such as relevance, usefulness, recommendation value and satisfaction. In the Recommender Systems Handbook, Ricci et al. [29] argues that utility represents the value that users receives in being recommended. As their own definition mentions, if the user enjoys the recommended items, he/she received useful recommendations. Moreover, utility has been defined as an order of preference of consumption. If users would only consume what they like most in the first place, therefore, recommending such items would help him/her find them easily, bringing usefulness to the recommendation Herlocker et al. [14]. Moreover, Adamopoulos et al. [1] cites the use of utility theory of economics for improving user satisfaction. Kotkov et al. [20, 21] also mention in a survey that utility or relevance relates to what the user is interested in consuming and therefore related to the tastes of the user.

As it can be seen, most of the definitions associate utility with the desires of consumption of the user and if the user enjoyed the recommendations. On such definition, metrics for assessing utility in recommendation should focus on how the user might react to the predictions made by a recommender. Ricci et al. [29] mention that utility could be measured by evaluating the rating that the user gives to predicted items after consuming them. This method is likely to be correct and capture if the recommendations brought value to the user, however it would involve in a costly online evaluation.

For offline evaluation, Herlocker et al. [14] mention the use of accuracy-based metrics for evaluating utility. The authors discuss the use of *predictive accuracy* metrics in order to evaluate if the user consumed the recommended items, usually in a train/test experiment. In this paper, we use the notation $util(R_u)$for utility, however there are many metrics for this purpose that the following subsections show.

### 2.1.1 Error metrics

Error metrics are widely used for predictive accuracy. *Mean Absolute Error* evaluates the difference between the ratings predicted by the recommender and given by the users [14]. Equation 1 show the MAE metric.

$$util(R_u) = MAE = \frac{\sum_{i \in R_u} p(i) - r(i)}{|R_u|} \quad (1)$$

Moreover, *Root Mean Squared Error* is another error metric, as it is shown in Eq. 2. Root Mean Squared Error calculates a larger difference for large errors in the rating prediction [29]. Both MAE and RMSE are calculated on the prediction list, therefore the metrics are divided by $R_u$. In addition, there are other error metrics, such as *Average RMSE, Average MAE* and *Mean Squared Error*.

$$util(R_u) = RMSE = \sqrt{\frac{\sum_{i \in R_u} (p(i) - r(i))^2}{|R_u|}} \quad (2)$$

#### 2.1.2 Precision and Recall

According to Ricci et al. [29], precision of a recommendation consists on the number of consumed (or rated) items in the recommendation list, as stated in the Eq. 3. Precision measures the rate of items in the recommendation list that the user likes and therefore consumed.

$$util(R_u) = precision = \frac{|C_u \cap R_u|}{|R_u|} \quad (3)$$

Recall, on the other hand, is calculated by the number of consumed items in the recommendation list out of the total number of items the user consumed [29]. Equation 4 show recall calculation. Authors have called precision and recall as *precision@N* and *recall@N*, where $N$ stands for the size of the recommendation list.

$$util(R_u) = recall = \frac{|C_u \cap R_u|}{|C_u|} \quad (4)$$

In applications, Zhang et al. [38] evaluated their recommender against *novelty, diversity, serendipity* and also used *rank* and *recall* in their metrics. Hurley and Zhang [15] also uses *precision* in their evaluations.

#### 2.1.3 ROC curves

Ricci et al. [29] mention the use of ROC curves in accuracy-based evaluation of recommendations. ROC curves measure the rate of items that the user likes in the recommendation list. Differently from error, precision and recall metrics, the calculation of ROC curves accentuate items that were suggested but the user disliked. Evaluation of algorithms in different scenarios could use the Area under the ROC curve (AUC) [29].

Herlocker et al. [14] also mention and exemplify that ROC curves could be plotted using the rate of useful and not useful items in a recommendation list. In this sense, a useful item could be defined if the user liked/consumed the item or not [14].

#### 2.1.4 Ranking score

Herlocker et al. [14] cites that *rank metrics* are useful in evaluating recommendations lists. Recommenders usually predicts ranked lists, however, users difficultly browse through all of the items. Therefore, ranking metrics could be interesting in measuring the utility and rank information altogether. One example is the *R-Score* metric which considers a deduction in the value of recommendations according to the rank position. Top ranked items are more valued rather then items in the tail of the metric [29].

Equation 5 show the *R-Score* metric, where $r(i,j)$ is the rating of the item $i$ in the rank, $d$ is a median rating and $\alpha$ represents a half-life decay value. Besides this R-score, there are other ranking scores metrics, such as *Kendall* and *Spearman* rank correlation and *Normalized Distance-based Performance Measure* [14].

$$util(R_u) = rank(R_u) = \sum_{j=1}^{|R_u|} \frac{\max(r(i_j) - d, 0)}{2^{\frac{j-1}{\alpha-1}}} \quad (5)$$

#### 2.1.5 Utility-based metrics for online evaluation

Utility is also evaluated with users in online experiments. In this sense, researchers usually make user experiments for testing the utility of their recommender systems or evaluate it when it is being applied in the industry. Such experiments are also a good way to measure the overall systems targets [32]. For these kinds of online experiments, some metrics are employed for evaluating the working recommender system.

Click-through-rate (CTR) is calculates the ratio of clicked/interacted recommended items out of the number of items recommended. It has been used since the early stages of the web in web/mobile advertisement and online marketing campaigns. CTR is also a major metric applied in the industry of recommender systems, as it helps to study how many items recommended to the users that they effectively consume. It has been mentioned or used in many work in the area, such as Farris et al. [10], Chu and Park [7] and Gomez-Uribe and Hunt [13]. The premise is that by clicking/interacting/consuming a recommended item, the user considers that recommendation useful. From a business point of view, it shows how effectively the recommender system is in predicting useful items to the user. The metric can be seen in Eq. 6.

$$util(R_u) = CTR = \frac{|C_u|}{|R_u|} \quad (6)$$

Retention is also a useful metric used in online evaluation of recommender systems [32] user utility and for business. Retention measures the impact of the recommender systems in keeping users consuming items or using the system. It has been applied in many scenarios, at it is has been a focus of evaluation is systems such as Netflix [13]. In an online monthly subscription fee based services, retention is an important business metric, evaluating retention of the recommendation is important to keep track of how long users will spend on their systems. Then, many algorithms try to predict items to maximize such metrics. In [13], online

**Table 2** Levels of novelty

| Level | Simplification | Description |
|---|---|---|
| Level 1 | Life level | An item is novel in the life of the user, that is, the user has never heard of the item in his/her life |
| Level 2 | System level | Item is unknown for the user according to the user's history consumption |
| Level 3 | Recommendation list level | Non-redundant items in the recommendation list |

retention experiments are performed as A/B tests with users and the retention delta is calculated, such as shown in Eq. 7. In the authors' research, retention is calculated as the difference between users in control ($p_c$) and test groups ($p_t$) in the A/B test the authors performed.

$$util(R_u) = \Delta_{retention} = p_t - p_c \tag{7}$$

Lastly, it is important to mention that the previously mentioned metrics for utility evaluation on recommender systems are applicable to online evaluation. For instance, accuracy-based metrics, such as error metrics, precision, recall are suitable to be used in online evaluation as well.

## 2.2 Novelty

The concept of novelty generally involves the idea of having novel items in the recommendation. Although it seems to be simple at first, novelty has various definitions in the literature. Therefore, in order to make the definition easier, we classify the novelty definitions and metrics into three levels, as it is presented in Table 2. Novelty metrics are called $nov(R_u)$ in this paper.

### 2.2.1 Life level novelty

There are some authors that define novelty in the life level. Kapoor et al. [19] described unknown items as never consumed or known in the users' lifetime. Ricci et al. [29] affirmed that novel items should be unknown to the user. The authors definition seems to be referring to the level 1 of novelty, as the authors further mention that a hypothetical way to measure novelty is to ask whether the users know the item. Additionally, Zhang et al. [38] considered that items consumed out of the influence of a RS should be considered by recommenders when issuing predictions. Then, items that the users have never known before in the users' life are novel items.

Creating metrics for measuring life level novelty is not trivial. A proper metric for level 1 of novelty would have to consider information out of the system's context in order to measure what the user knows and do not know. No metrics surveyed seem to be evaluating life level of novelty.

### 2.2.2 System level novelty

The system level novelty has many definitions in the literature. In a simplified way, a novel item for a user is one that the user has none or little knowledge about [12]. Herlocker et al. [14], Iaquinta et al. [16, 17] state that novelty is when a RS predicts items that the user does not know about and might not discover by other sources. Moreover, novelty has also been defined as how different the recommended item is when compared towards what the user has consumed [36]. Lastly, novelty has also been defined as the proportion of unknown items in the prediction list for the user [15]. In practice, these definitions would only consider novel items when observing *previously consumed items* in the history of consumption of the users; items consumed outside of the system are not taken into consideration. In summary, even though the authors use different words, they still have the same meaning: level 2 of novelty means items that the user does not know, when considering the system's information.

Most metrics proposed for evaluating novelty in the literature fit in the level 2. Nakatsuji et al. [26] propose a metric that calculate novelty in recommendation list as the similarity between the items in the recommendation list and in the history of the user ($H_u$). The metric is shown in Eq. 7. The authors use the classes of items for measuring the distance between items. $d$ is a distance function and $class(i)$ represents the classes of item $i$. This idea can be extrapolated for features or genres of items. A summarization of the metric is shown in Kotkov et al. [20].

$$nov(R_u) = \sum_{i \in R_u} min_{j \in H_u} d(class(i), class(j)) \tag{8}$$

The metric proposed by Zhou et al. [40] and used by Zhang et al. [38] calculates the sum of the popularity of the items in the recommendation list of the user. Equation 8 show the novelty metric. The popularity ($pop$) of an item can be calculated, for instance, by the number of users that consumed it. While definition of novelty made by Zhang et al. [38] can be considered level 1, the authors' metric is related to level 2, since the popularity is calculated on the amount of users that consumed the item, using the users' consumption data. Therefore, the novelty of the

items is still in system level. Popularity based metrics for novelty proposed in Ricci et al. [29] and surveyed by Kotkov et al. [20, 21] have similar behaviour. As Eq. 8 shows, the metric simply calculates novelty of a recommendation lists, by calculating the popularity of items in the list. The authors also provide variations of the metric, such as $-log_2 \frac{pop(i)}{|U|}$, which is similar to the metric used in Zhang et al. [38].

$$nov(R_u) = \sum_{i \in R_u} \frac{log_2 pop(i)}{|R_u|} \tag{9}$$

$$nov\big(R_u\big) = 1 - \frac{|pop(i)|}{|U|} \tag{10}$$

#### 2.2.3 Recommendation list level novelty

Level 3 involves novelty in the recommendation list level, that is, items not repeatedly recommended. In this sense, novelty is defined as not repeated items in the recommendation list, not involving users' information. Adamopoulos et al. [1] said that novelty is related to non-redundant items in recommendation lists that the user does not know about. In short, level 3 is the extreme case of level 2, where not even redundant items in the recommendation list or repeated recommendations are allowed.

Metrics for measuring level 3 of novelty solely involve investigating items in the recommendation lists. No users' information is required in metrics for level 3 of novelty. In this sense, Eq. 10 calculates the similarity of items in a recommendation list [4]. Again, $d(i,j)$ means the distance between items $i$ and $j$. However, the metric looks like a metric for intra-list similarity and may not be measuring novelty.

$$nov(R_u) = \frac{1}{|R_u| - 1} \sum_{j \in R_u} 1 - d(i,j) \tag{11}$$

Also, Vargas and Castells [36] proposed a distinct metric for measuring novelty in recommendation list. Equation 11 shows their proposed metric. The metric takes into consideration the position of the items in the ranked recommendation list for calculating a discount ($disc(i_k)$) of browsing through the list. Moreover, the metric also calculates the probability of the user has seen the item ($p(seen|i_k)$) while browsing. Since this probability may or may not consider the users' consumption information, this metric is best classified between level 2 and 3 of novelty.

$$nov(R_u) = \sum_{k=1}^{|R_u|} disc(k)(1 - p(seen|i_k)) \tag{12}$$

### 2.3 Diversity

Diversity is a concept concerned with the diversity of items in the recommendation list. It has also been widely studied by previous researchers. For diversity metrics, the notation used in this paper is $div(R_u)$.

According to the Ricci et al. [29], diversity in RSs have the contrary effect of similarity. The authors state that recommendation lists with low variety may not be of interest of the user. Moreover, one of the earliest work concerned with diversification in recommendation is [41]. The authors argue that the RSs usually predict similar items compared to the user's consumption history. Therefore, diversity means balancing recommendation lists to cover the user's whole set of interests [41]. In addition, Vargas and Castells [36] state that diversity refers to the variety of the items in the recommendation list. Moreover, Hurley and Zhang [15] and Zhang et al. [38] reinforce the definition of diversity from [41] stating that it is related to the variation of items in predictions of a RS.

Differently from novelty, the definitions of diversity are largely consistent in the literature. All the authors surveyed in this work agree that diversity represents variety of items in recommendation lists.

As a result of this definition, the proposed metrics tend to calculate diversity as a dissimilarity between the items in the recommendation list. Ziegler et al. [41] proposed a metric for intra-list similarity, as Eq. 12 show. The function $d(i,j)$ calculates the distance between items $i$ and $j$ in the recommendation list $R_u$. This metric actually captures the similarity of the list; therefore, low values for this metric represent a more similar list, in which the items are similar to one another.

$$div(R_u) = \sum_{i \in R_u} \sum_{j \in R_u} \sum_{i \neq j} d(i,j) \tag{13}$$

The intra-list similarity metric was also used by other works in diversity. Zhang et al. [38] used the metric proposed by Ziegler et al. [41] and chose the cosine similarity as the distance function. The metric can be seen in Eq. 13. Moreover, Hurley and Zhang [15] used a similar diversity metric as 11, where the distance function ($d(i,j)$) is calculated by a Collaborative Filtering memory-based similarity metric.

$$div(R_u) = \sum_{i \in R_u} \sum_{j \in R_u} \sum_{i \neq j} cossim(i,j) \tag{14}$$

Individually, Vargas and Castells [36] proposed a distinct metric for calculating similarity. Their metric, as stated in Eq. 14, is a more specific case of the intra-list similarity. The metric takes into consideration a relative rank discount function for the position of each pair of items being analyzed

($disc(k)$ and $disc(l|k)$). Moreover, the metric also uses a distance function ($d(i_k, i_l)$) between the items, for instance cosine similarity distance. The approach proposed by Vargas and Castells [36] resembles techniques used in Rank Information Retrieval for both the authors' novelty and diversity metrics.

$$div(R_u) = \sum_{k=1}^{|R_u|} \sum_{l=1}^{|R_u|} disc(k) disc(l|k) d\left(i_k, i_l\right) \; \forall i_k \neq i_l \tag{15}$$

## 2.4 Unexpectedness

Unexpectedness is a concept that has been increasingly mentioned in the literature, but it is still involves uncertain definitions. It is usually linked to surprise and avoidance of obviousness in recommendation. The notation used to show the unexpectedness metrics in this paper is $unexp(R_u)$.

Unexpectedness was firstly stated a component of serendipity by McNee et al. [24] and Ge et al. [12]. In both researches, the authors use the term *unexpectedness* to define the idea of *surprise* in recommendation. Moreover, Kaminskas and Bridge [18] also mention that unexpectedness represents surprise in recommendation.

Unexpectedness has also been defined as a *divergence from expected recommendations*. In Murakami et al. [25], the authors also state unexpectedness as a part of serendipity and describe it as a deviation from items that the user expect to consume, however the authors mostly focus on serendipity. Adamopoulos et al. [1] also explain that serendipity and unexpectedness concepts have been overlapping. The user expectations consist on the set of items the user would like to consume next or the items the user forecast to be recommended [1]. Therefore, unexpectedness would be a deviation from these expected items, evading obvious and uninteresting recommendations, with the possibility of surprising the user [1].

Measuring unexpectedness is not trivial, due to its overlapping definitions. Two set of metrics have been proposed in the literature: metrics based on a primitive recommender and metrics based on principles not involving a primitive method. We present both set of metrics as follows.

### 2.4.1 Primitive recommender based unexpectedness

According to Ge et al. [12], a primitive recommender usually predicts items that the user expects to consume. Such consideration is reasonable, considering unexpectedness as a deviation from expected recommendations. Therefore, Eq. 15 present unexpectedness as the items in a recommendation list ($R_u$), but not in a set of prediction made by a primitive recommender ($PM_u$), proposed by Ge et al. [12].

$$unexp\left(R_u\right) = R_u - PM_u \tag{16}$$

The primitive recommender idea was later enhanced by Adamopoulos et al. [1], where the authors measure the rate of unexpected items in a recommendation list ($R_u$), such as shown in Eq. 16. In this metric, $E_u$ is the set of expected items for the user. In short, $E_u$is the same as $PM_u$.

$$unexp\left(R_u\right) = \frac{R_u - PM_u}{|R_u|} \tag{17}$$

The problem with primitive recommender based metrics lies in choosing an appropriate primitive recommender. The choice should be made considering the recommendation's context. Users may have different expectations for movies and songs, for example. Moreover, different primitive recommenders will lead to different unexpectedness values. Therefore, using a primitive recommender may not be a trivial way to measure unexpectedness.

### 2.4.2 Non primitive recommender based unexpectedness

Metrics to assess unexpectedness based on principles not involving a primitive recommender also exist in the literature. Kaminskas and Bridge [18] attempt to calculate surprise using a metric represented by Eqs. 17 and 18. The Point-wise mutual information function ($PMI(i,j)$) calculates the probability of two items $i$ and $j$ be rated by the users. The PMI function is $PMI(i,j) = \frac{log_2 \frac{p(i,j)}{p(i)p(j)}}{-log_2 p(i,j)}$, where $p(i)$ is the probability of item $i$to be rated by users. In this case, this metric is comparing the recommended items and the history of the user, checking if the user is likely to know the predictions. Nevertheless, the metric may not be effectively measuring whether the user gets surprised with the recommendations. Besides, as the authors explain, *PMI* function may be biased towards rare items, which may always be considered unexpected to the user.

$$unexp\left(R_u\right) = \sum_{i \in R_u} \sum_{j \in H_u} PMI(i,j) \tag{18}$$

$$unexp\left(R_u\right) = \sum_{i \in R_u} max_{j \in R_u} PMI(i,j) \tag{19}$$

Akiyama et al. [3] proposed an unpersonalized metric for unexpectedness that does not consider the users' information. The metric, as it is shown by Eq. 19 use an idea of co-occurrence, but it is limited to items and their features. For instance, $I_v$ calculates the number of items that have feature $v$ and $I_{v,w}$ calculates the number of items that have both features $v$ and $w$. The probability of co-occurrence uses items' features to measure how similar these items are. The author explains how to calculate the unexpectedness of a single item, however, one could calculate unexpectedness for entire recommendation list $R_u$ by summing or averaging the unexpectedness of an item. Since this metric is not personalized, it is unlikely that it is measuring unexpectedness to users.

$$unexp(R_u) = \frac{1}{\frac{1}{|F_i|}\sum_{v,w \epsilon F_i} \frac{I_v}{I_V + I_w - I_{v,w}}} \quad (20)$$

In summary, it can be seen that definitions and metrics for unexpectedness are unclear in the literature. In general, unexpectedness means surprise and avoiding expectations of users. However, the definitions presented somewhat overlap with other concepts such as serendipity, and there are different metrics to measure unexpectedness. In Silveira et al. [34], the authors summarize metrics for unexpectedness and propose an evaluation methodology for unexpectedness evaluation in recommender system.

## 2.5 Serendipity

Serendipity has been increasingly used in recommender systems, however it has a complicated definition. In this section, metrics for serendipity use the notation $ser(R_u)$.

The term serendipity means a *lucky finding* or a *satisfying surprise*. According to Ricci et al. [29], serendipity represent surprising recommendations. One of the earliest mentions of serendipity in the literature comes from [14]. The authors use the word serendipity as a concept of surprising and interesting item for the user. The same is stated by Iaquinta et al. [17], where they mention serendipity represent items that the users would difficultly find. Moreover, Ge et al. [12] define serendipitous items as surprising and pleasant. As it can be seen, most authors agree that the serendipity concept involve a *good and pleasant surprise*. However, it is necessary to state that serendipity is a perception of the users with regard to the recommendations they receive [12, 20].

Other definitions of serendipity can be found in the literature. For instance, Zhang et al. [38] and Kotkov et al. [20, 21] say that serendipitous recommendations are *unusual and surprising* to the users. Furthermore, serendipity can also be seen as *good emotional* answer to a novel recommendation that the users were not expecting to receive [1]. In this definition, Adamopoulos et al. [1] conclude that serendipitous recommendations are novel, unexpected and useful. Ge et al. [12] and Murakami et al. [25] also associate serendipity and unexpectedness.

Several other work have been studying serendipity problem in recommendation. For example, Lu et al. [11], Onuma et al. [22] and Gemmis [27] are some works that try to propose algorithms for serendipity. Additionally, Kotkov et al. [20, 21] made a large survey on serendipity. In summary, it can be concluded that even though serendipity has a hard to understand definition, most authors agree that it represents a *delightful surprise* and provide useful and surprising items to the user.

Metrics have been proposed to measure serendipity in recommendation lists and most of them have some relation to the concepts that serendipity is involved to: level 2 of novelty, unexpectedness and utility. Some metrics attempt to use unexpectedness notion of a primitive recommender. Therefore, we divide the metrics into the primitive recommender based metrics and non primitive recommender based.

### 2.5.1 Primitive recommender based serendipity

To our knowledge, the first metric proposed to evaluate serendipity was presented in Murakami et al. [25]. The metric can be seen in Eq. 20. $PM_u$ is the primitive recommender. Moreover, the metric also uses a relevance function ($rel$), which calculates if the predicted items are relevant to the user or not 0 for relevant or 1 for irrelevant. Moreover, the position in the recommendation rank is also taken into consideration ($\frac{count_u(k)}{k}$).

$$ser(R_u) = \sum_{k=1}^{|R_u|} \max(R_u[k] - PM_u[k], 0) rel(i_k) \frac{count_k(k)}{k} \quad (21)$$

The metric proposed by Murakami et al. [25] was also used by Ge et al. [12] and it can be seen in Eq. 21. The metric was simplified and the rank of the items in the list were not considered. Moreover, $UNEXP_u$ represent the surprising items for the user $u$, which is calculated as an unexpectedness metric ($UNEXP_u = R_u - PM_u$). The author maintains the utility function.

$$ser(R_u) = \frac{\sum_{i \epsilon UNEXP_u} \text{utility(i)}}{|R_u|} \quad (22)$$

Adamopoulos et al. [1] also utilizes the same metric with some terminology variations, as it can be seen in the Eq. 22. For the authors, serendipity is said to be the rate of not expected ($R_u - E_u$). Again, $E_u$ represents the set of expected items and can be replaced as $R_u$. $USEFUL_u$ is the set of useful items in the recommendation list.

$$ser(R_u) = \frac{(R_u - E_u) \cap USEFUL_u}{|R_u|} \quad (23)$$

As mentioned in the unexpectedness subsection, primitive recommendation based metrics are dependent on choosing a primitive recommender. Selecting different primitive recommenders will result in different values of serendipity. For instance, in Adamopoulos et al. [1], for selecting expected items for the user, the authors use the profile of the users and a set of rules about the two datasets evaluated. Moreover, a utility function must be appropriately selected for calculating the usefulness of the items to the user. Again, different utility functions will result in different values for serendipity.

#### 2.5.2 Non primitive recommender based serendipity

Zhang et al. [38] proposed a metric that calculates the cosine similarity between recommended items ($R_u$) and the history of consumption of the user ($H_u$). The metric is shown in Eq. 23. In this case, low values represent more serendipitous recommender lists. The metric is reasonable, since the recommended items should not be very similar to the user's consumption profile. However, the metric does not evaluate the usefulness of the recommendations, only the element of surprise is considered. Therefore, it may be evaluating novelty or unexpectedness, instead.

$$ser(R_u) = \frac{1}{|H_u|} \sum_{i \epsilon H_u} \sum_{j \epsilon R_u} \frac{cossim(i,j)}{|R_u|} \tag{24}$$

In short, serendipity is a complex concept. Even though most authors agree that it represents usefulness and surprise, it is not known if the metrics are effectively evaluating serendipity. Metrics, such as Murakami et al. [25] ones, are sensitive to the selection of a primitive recommender. Moreover, the literature has distinct metrics to evaluate the same concept.

### 2.6 Coverage

Coverage is another concept that has been analyzed by previous researches in recommender systems. Although there have been few studies proposing metrics for coverage, it still worthy to mention due to its potential relation with the other explored concepts in this paper. In addition, other concepts such as trust, risk, robustness are far less studied than coverage. The notation of coverage used in this paper is *cov*. Coverage evaluates the whole RS, not a recommendation list. Moreover, three kinds of coverage are mentioned in the literature: item space coverage and user space coverage, as presented by Ricci et al. [29], and genre space coverage, as proposed by Vargas et al. [35]. In our survey, only metrics for item space coverage were found in the literature.

#### 2.6.1 Item space coverage

Item space coverage refer to the extent of items that a recommender system is able to make predictions. Herlocker et al. [14] described that the coverage of a recommender system consists on set of items that the system is capable of working with. A recommender with lower item coverage limits the recommendations for the user. Low coverage prevents them from finding useful items to consume, impacting in the users' satisfaction and in the overall sales of the system [14]. Ricci et al. [29] says that item space coverage refers to the ratio of items effectively being recommended to users, such as mentioned by Herlocker et al. [14].

Moreover, Ge et al. [12] also define coverage in their work. The author's definition is the same as the item space coverage: proportion of items that the system is able to predict. The authors also divide coverage into prediction coverage and catalog coverage. While prediction coverage means the set of items in the system that can be predicted to users, catalog coverage refers to the coverage in prediction lists.

Few metrics are proposed for coverage, since it is not widely used in practice. Ge et al. [12] proposes two metrics for calculating prediction coverage (Eq. 24) and catalog coverage (Eq. 25). The prediction coverage metrics is simply the rate of items for which prediction is possible ($I_p$) over the size of the item set. For catalog coverage, Eq. 25 shows the rate of distinct items recommended over a period of time to the user [12]. In both metrics, coverage seems to capture the proportion of items that the system is able to work with.

$$cov = \frac{|I_p|}{I} \tag{25}$$

$$cov = \frac{|\sum_{j=1..n} I_L^j|}{|I|} \tag{26}$$

Ricci et al. [29] also show a metric for catalog coverage derived from sales diversity. Equation 26 shows that the coverage concept is measured by $p(i)$, which represent the number of users that chose the item $i$. The metric seems to capturing coverage and how is the disparity of recommendation of distinct items.

$$cov = \frac{1}{|I| - 1} \sum_{j=1}^{|I|} (2j - |I| - 1) p(I_j) \tag{27}$$

#### 2.6.2 User space coverage

User space coverage refers to the proportion of users that a RS can predict items to. According to Ricci et al. [29], in certain kinds of recommendation problems, the predictor may not have high confidence of the accuracy of the prediction for the users. Therefore, user space coverage would measure the rate of users who receive effective recommendations. The author further mentions about the cold start problem, where the confidence is low for new items and users in the system [29]. No metrics for user space coverage were found.

#### 2.6.3 Genre space coverage

Lastly, Vargas et al. [35] proposed a different kind of coverage, which is genre coverage. In their work, the authors study genre coverage, redundancy and size-awareness. Specifically, for genre coverage, they define as the number of

distinct genres of items that are effectively recommended to users [35]. In this sense, the author's definition of coverage is more related to diversity of the recommendation lists.

### 2.7 Performances of state-of-art Recommender Systems in terms of different metrics

In order to further enhance the discussion on how good the Recommender Systems are and their evaluation, we present a review on work in the literature that performed evaluations on state-of-art recommenders regarding the aforementioned concepts. This brief review does not focus on comparisons, since different work decide on evaluating their recommender system regarding different concepts. Furthermore, even same concepts are evaluated by different metrics. Therefore, we center our attention on summarizing the main achievements in performance with regard to the six concepts of evaluation. For the sake of summarization, we included Table 4 in the appendix summarizing the performances of the State-of-Art recommenders reviewed in this subsection.

With regard to **utility**, most articles evaluate their recommender systems with offline accuracy metrics. As an example, some collaborative filtering algorithms were initially evaluated through MAE and RMSE metrics. Wen [37] reviewed Item-based KNN, Item-based EM and Sparse SVD for the recommender system problem. The author evaluated the RMSE metric (Eq. 2) on a Netflix database[1] and obtained results as high as 0.95, 0.91 and 0.8974 for the test set for the Item-KNN, Item-EM and SVD methods, respectively.

**Novelty** has been evaluated in many ways in the literature, however using different metrics and definitions. For instance, Vargas and Castells [36] evaluated state-of-art methods using level 3 novelty metric stated in Eq. 10. The authors used Matrix Factorization, IA-select and MMR methods in their evaluation of novelty and considered a relevance discount, which resulted values of novelty of 0.058, 0.0639, 0.0620, respectively, in the Movie Lens[2] dataset. Not considering the relevance discount, the authors achieved results on novelty of 0.76, 0.8080 and 0.7605, respectively, also for the Movie Lens dataset. The authors also evaluate the same methods on Last.fm[3] dataset, where they obtained 0.2671, 0.3462 and 0.2439 of novelty ratio, respectively, considering the relevance discount; and the authors obtained 0.8949, 0.8912 and 0.9133, respectively, disregarding the relevance discount. It can be seen that more novelty was achieved in the Last.fm dataset. However, it is noteworthy that, in this case, high novelty may mean low values of accuracy or other utility measurements, because novel items not necessary are highly useful to the users.

Vargas and Castells [36] also evaluated the state-of-art recommenders with regard to their **diversity** metric. In this case, the authors used the diversity metric described in Eq. 13. For Movie Lens dataset, the recommenders Matrix Factorization, IA-select and MMR achieved 0.0471, 0.0537 and 0.0510, respectively, considering a relevance discount. Ignoring the relevance, the resulting diversity ratio was elevated to 0.7164, 0.8289, 0.7191. Again, similar to the novelty statement above, high diversity in this case may negatively impact the utility. In addition, De Pessemier et al. [8] evaluated Group Recommendation on Movie Lens dataset and diversity was evaluated for User based and Item Based Collaborative Filtering algorithms and an SVD method, although the authors do not further specify which algorithm was used. The metric used for diversity was the intra-list similarity, as Eq. 11 states. When considering the group recommendation size equal to 1, the diversity similarity was 0.7, 0.81 and 0.64 for User-based CF, Item-based CF and SVD methods.

Regarding the concept of **unexpectedness**, it is not usually assessed by work in the area in the literature. Adamopolous et al. [1] is one of the few work in the literature to evaluate the unexpectedness and compare their own method with state-of-art baselines. The authors, evaluated Item and User KNN and Matrix Factorization methods. However, in this work, the authors evaluate many combinations of parameters and they do not mention which baselines are used in comparison with their recommending methods, instead they calculated the average values of their experimental settings. The authors evaluate unexpectedness through the metric shown in Eq. 15. Two datasets were analyzed in this case: Movie Lens and Book Crossing[4]. For Movie Lens dataset, the unexpectedness ratio for their baseline metrics are 0.71 and 0.75 for recommendation lists size 10 and 100 respectively, using related movies as expected recommendations. For the Book Crossing dataset, the values of unexpectedness are more modest, where it is 0.3 and 0.38 for size 10 and 100, respectively, using related books as expected recommendations.

Although there are work which study the **serendipity** concept and propose new methods and metrics, few of them effectively use state-of-art baseline as comparison. Lu et al. [22] evaluated serendipity on two datasets: Netflix and Yahoo! Music[5], using popular methods: SVD, SVD++

[1] Dataset retrieved from Netflix Prize at http://www.netflixprize.com.

[2] Movie Lens Dataset from Group Lens available at http://grouplens.org/datasets/movielens/.

[3] Last.fm dataset provided by Celma and Herrera [5].

[4] Book Crossing dataset made available by Ziegler et al. [41].

[5] Yahoo! Music dataset used was made available by Dror et al. [9].

and SVDNbr. The metric used for assessing serendipity was analogous to Eq. 21. The authors used different loss functions and optimizing methods, however their best serendipity results for each method for the Netflix dataset was 0.2534, 0.3036 and 0.2978 for SVD, SVD++ and SVDNbr, respectively. For the Yahoo! Music dataset, the best results were: 0.1995, 0.2771 and 0.3834 for the three methods, respectively.

Also, there has been attempts to evaluate a temporally evolving system with regard to the aforementioned concepts. Shi et al. [33] addressed the issue of the performances of recommender systems in a temporally dynamic system. The authors described the datasets (Movie Lens and Netflix) as bipartite networks of users and items divided into a series of subsets which considers the recommendations and the time series. Authors used a recommendation based evolution method to simulate the temporal dynamic of three common collaborative filtering strategies. They evaluate their results at each time step with utility (RMSE Eq. 2), intra-list similarity (similar to Eq. 13), system-level novelty (metric similar to Eq. 9). For Movie Lens dataset for instance, the authors showed to be able to maintain low levels of similarity, around 0.025 with Item-based CF in the temporal evolving process.

Lastly, as mentioned earlier, the **coverage** concept is also under-explored, both in metrics definitions and work concerning this concept. Nevertheless, one of the few authors that evaluate coverage is Adamopolous et al. [1]. In the authors' work, catalog coverage is evaluated using metric similar to Eq. 24. For Movie Lens dataset, the authors obtained 0.05 and 0.15 rate of items ever recommended to the user for the size of the prediction list of 10 and 100, respectively. Moreover, for the Book Crossing dataset, the coverage ratio is about 0.2 and 0.5 for 10 and 100 sizes of the prediction list.

It is important to notice that the reviewed work consider different datasets, methods, evaluation metrics, implementations and evaluation methodologies when studying the impact of the user satisfaction concepts in recommendation. We emphasize the need of a novel work which can effectively provide **user satisfaction** evaluations for the state-of-art recommenders under the same evaluation scenario and datasets considering the mentioned concepts. Moreover, in order to complement such research, such work could also analyse the impacts of attempting to evaluate and optimize all of those concepts simultaneously and attempt to combine recommendations with different objectives, similar to what was performed by Zhang et al. [38] and Ribeiro et al. [28] in a limited level.

## 3 Categorization and relationships among the concepts

According to the main definitions and metrics existing the literature about the concepts of evaluations in recommendation, it is noteworthy to categorize and establish relationships among them. In this section, the concepts are categorized regarding the user dependence and are inter-related with one another.

### 3.1 User dependency properties of the concepts

The metrics presented before have the main objective to evaluate how good the recommender system is in making predictions for the user under different perspectives, evaluating whether the predictions are novel, diverse, surprising and useful. Although there are many metrics and definitions, they can be classified regarding the property of user dependency. Some metrics are sensitive to users' information, while others do not depend on the users and simply evaluate the system or recommendation list. Therefore, the metrics can be classified in *user-dependent* and *user-independent*.

#### 3.1.1 User-dependent concepts

Metrics for user-dependent concepts require user information. They are usually calculated by comparing the items recommended and history of consumption of the user. *Utility, life and system level novelty, unexpectedness* and *serendipity* are user-dependent concepts.

The accuracy-based metrics for *utility* requires user consumption data to verify if the user enjoyed the recommendations made or not, no matter in an online or offline experiment. For instance, *RMSE, MAE, precision* and *recall* metrics requires the predictions for the users. Then, utility is a user-dependent metric.

In addition, as *life and system level novelty* represent unknown items for the user, metrics for evaluating them require the users' information. *Life level novelty* is defined in items that is novel to user in his/her life. Even though there is no metric for this level of novelty, possible future metrics would very likely require user information. Therefore, level 1 of novelty is user-dependent. Regarding the *system level novelty*, it is clear to note the metrics are user-dependent (Eqs. 6–8) because these metrics require consumption information from the users, in order to verify whether the recommendations are novel.

Furthermore, *unexpectedness* is also a user-dependent concept, since it considers the expectations of users, which involves, the users' profile. Most of the presented metrics for unexpectedness (Eqs. 14–16) need user information and predictions made by other recommenders.

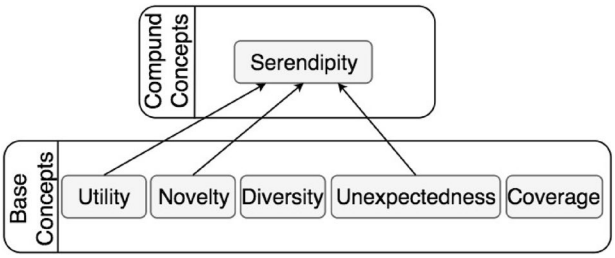


**Fig. 2** Relationships of the evaluation concepts according to previous work

Since *novelty, unexpectedness* and *utility* are user-dependent, *serendipity*, which is based on these three concepts, is also a user-dependent concept. Most of the metrics surveyed for evaluating *serendipity* require user information, such as the history of the user, either based on a primitive recommendation or not (Eqs. 19–22).

#### 3.1.2 User-independent concepts

Metrics for evaluating user-independent concepts do not require user information. Concepts classified in category evaluate other areas rather than users' rating or their history of consumption. In this category lies *diversity, recommendation list level of novelty* and *coverage*.

*Diversity* is considered a user-independent concept as it exclusively assesses the variety of items in a recommendation list. Diversity metrics (Eqs. 11–13) only require the recommendation list itself and information about the items, therefore it does not depend on user information.

Similarly, metrics for *level 3 of novelty* are only computed using the recommendation list items information. Data about users' consumption is not required in this level as it only measures non-repeated items in the list. Therefore, recommendation list novelty is also user-independent.

Lastly, *coverage* is a concept that evaluates the system behaviour. Metrics proposed for item space coverage only use the items information. Users' information about consumption is not required. Then, coverage can also be categorized as a user-independent concept.

### 3.2 Relationships among the concepts

According to the survey made and the previous discussion, the concepts can be organized into categories that represent their relationship with regard to each other: base and compound concepts. Regardless of the categorization on user dependence, some concepts are related to each other, considering that their definitions take others into consideration. Figure 2 shows the current relationships among the concepts. The arrow represents a relationship between the concepts as it has been mentioned in the literature.

There are five base concepts: *utility, novelty, diversity, unexpectedness* and *coverage*. These concepts are grouped together as their definitions on the literature do not depend on any other concept. Moreover, base concepts are used in the definitions of compound concepts. Currently, only utility, novelty and unexpectedness are related with compound concepts. Diversity and coverage are currently isolated in this relationship scheme. So far, no definitions and metrics for diversity and coverage mention other concepts.

Individually for *unexpectedness*, even though some definitions mention novelty [1], the authors' usually state that surprising recommendations can be either novel or not. Therefore, since there is no limitation for unexpectedness regarding novelty, as there is for serendipity, it can be currently stated as a base concept.

The only current compound concept is serendipity. The definition of serendipity consists on three components: *utility, novelty* and *unexpectedness*, so that serendipitous recommendations can be called a *delightful surprise* [1]. Herlocker et al. [14] and Ge et al. [12] say that serendipitous recommendations are novel by default, as the user must not have knowledge about them. Moreover, they need to be unexpected as it is the surprise part of the definition [1]. Lastly, since it involves a positive response, there is a direct relation between serendipity and how useful the prediction is for the user [1, 12]. Therefore, in this classification, serendipity is a compound concept.

The currently relationships of the evaluation concepts, as stated by previous work, may be oversimplified. There might exist relationships among the concepts' definitions and underlying correlations among them. We cover these potential relationships in the next section.

## 4 Potential research topics on evaluation

There are still many open issues with evaluations of the presented concepts and metrics. Therefore, we present potential research topics on concepts of evaluation in recommendation. We re-examine the relationship between the concepts and discuss the factors related to users' satisfaction. Lastly, we also discuss about online evaluations in recommendation.

### 4.1 Re-examining the intentions of the concepts and their relationships

As it was pointed out earlier, there is still much open issues with metrics in recommendation. Metrics for utility concept are an example. One of the reasons researchers started to investigate other concepts rather than predictive accuracy in RS is that it is not known whether accuracy really represents user satisfaction [14]. Currently, a set of concepts is used in evaluating whether the users enjoy the predicted

items. Therefore, future researches could further study utility metrics in order to check if evaluating utility is equal to predictive accuracy.

Moreover, we emphasize the classification of *novelty* in three levels. The existing metrics only explore level 2 and 3 of novelty. Evaluation about the first level of novelty is difficult and so far there are no metrics. Then, more research is required about level 1 of novelty. Future metrics for *life level of novelty* need to consider information out of the system in order to model the users' lifetime consumption. It is clearly a challenging task.

*Diversity* seems to be related to Information Retrieval diversity, by analyzing the metrics, such as the one proposed by Vargas and Castells [36]. Information retrieval techniques may have a lot to contribute to diversity metrics in RSs. Moreover, *diversity* can be correlated to *unexpectedness* and *serendipity*. Attempting to recommend surprising items may enhance diversity in recommendation. Increasing unexpectedness implies recommending items not usually recommended in expected recommendations. Therefore, the variety of items may be enhanced by issuing unexpected recommendation lists. Recommending serendipitous items might result in the same effect, since unexpectedness is one its components.

*Serendipity* and *Unexpectedness* seem to be the new directions of evaluations in recommender system. However, as surveyed, definitions and metrics remain overlapping and somewhat fuzzy in the literature. Since there are many metrics to evaluate the two concepts, new researches in the area do not know which metric to use in evaluation. Therefore, more research is necessary to clearly separate the definitions and properly establish metrics for unexpectedness and serendipity. Further, the two definitions for both concepts have limitations. *Unexpectedness* can reduce tediousness and obviousness and it is not limited to novel recommendations. *Serendipity* means surprising and useful, but the definition only allows novel items [17]. This is a limiting factor for applying the concept in real-world problems. A scenario requiring evaluation on surprising, useful and not necessarily novel could exist. For instance, forgotten items or items consumed far in the remote past could become serendipitous in some moment in time, but these items are not necessarily novel, because the user knows them, according to the level 2 of novelty.

In this sense, future researchers could discuss relaxing *serendipity's* definition, so that it is not limited to novel items. Otherwise, they could enhance *unexpectedness'* definition to include utility as well. Alternatively, a new concept could be proposed, which could simply be composed of surprise and usefulness, but no restriction with regard to novelty.

*Coverage* is an under-explored concept. Few work attempted to evaluate item space coverage in practice,

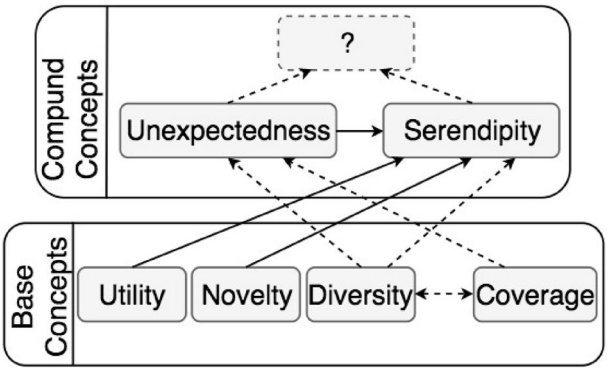


**Fig. 3** Potential correlations among the evaluation concepts

because the metric does not evaluate if the user enjoy the recommendation or not, instead it evaluates the RS. Future works could research the applicability of coverage in real scenarios. Furthermore, item space coverage could be related to other concepts. For instance, coverage could be correlated to *recall*, as recall is related to items not recommended. Moreover, uncovered items may enhance *unexpectedness* to the user. Items not effectively recommended in the system could be unexpected, since the user do not reach them. Lastly, there could be a correlation between *coverage* and *diversity*. Increasing the system coverage would increase the rate of items that are recommended; therefore, items not commonly recommended would be issued in recommendation lists, potentially decreasing the intra-list similarity among the items. Coverage, as well as its potential relationships with other concepts, is an open under-explored issue, which could be inspected in future research.

According to the ongoing discussion, Fig. 3 shows the discussed changes in the relationships among the concepts. Dashed arrows present the possible correlations among the concepts, which has not been aware by previous work, as previously described in this session. The figure has dashed arrows for representing the possible correlation between diversity and coverage, and between diversity and serendipity. *Unexpectedness* is transferred to the compound concepts group in this scheme, because it would have relationship with *coverage* and *diversity*. Moreover, the hypothetical new concept combining the desired features from serendipity and unexpectedness is also represented in Fig. 3.

### 4.2 A new direction: user satisfaction

In order to better research concepts and metrics for evaluation in recommendation, it is important to consider what are the factors that compose users' satisfaction. Ultimately, RSs have the objective of satisfying the users' interest of consumption. In offline and mainly in online experiments, it is important to study what are the

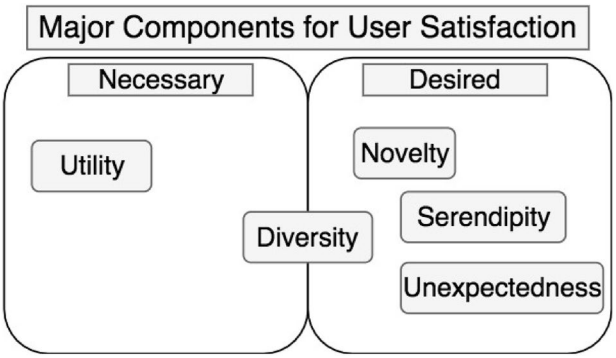


**Fig. 4** Necessary and desired components for user satisfaction

interests of the user when using a system, so that concepts and metrics can be designed to evaluate recommenders. Recommenders can also be designed to fulfill the users' interest.

Users have an intrinsic need for useful recommendations, but sometimes, users also desire to have new experiences and get surprised. Figure 4 shows user satisfaction requirements sorted from the needs to the desires of the user. *Needs* represent what the recommendations must have, while the *desires* represent what the recommendations could also have. *Utility* is the main need the users have, since they want recommender systems to suggest useful items according to their tastes. Even though they are important, *serendipity* and *unexpectedness* represent desires for surprise in recommendation. *Diversity* places in the border of the two constraints, because users desire and need some variety in recommendation. Moreover, *novel items* are not useful to the users, instead novelty is a concept that users would desire to be suggested from time to time, in order to extend their tastes. Lastly, *coverage* is not evaluated this image, since it is nor a need neither a desire for the user.

In short, user satisfaction should be considered when modelling new recommenders. Future researches could evaluate prediction lists and their algorithms using all the concepts presented. Offline and online experiments are important, even though the experiments with users can be costly to perform. Moreover, future works could consider designing algorithms that simultaneously balance the surveyed concepts in a single recommendation list. Zhang et al. [38] attempted to balance novelty, diversity and serendipity, while Ribeiro et al. [28] attempted improving accuracy, novelty and diversity, accordingly. Nevertheless, novel attempts in studying and balancing all of the mentioned concepts are necessary for increasing user satisfaction considering the needs and desires of the user.

### 4.3 Methodology of online evaluation

Most of the work surveyed in this paper refer to offline experiments using the concepts for evaluations of recommendation. Experiments usually elaborate train and tests sets on a dataset and perform evaluations by comparing the recommendations with what was consumed in the test set. Since this approach does not always correctly model the consumption behaviour of the users, as an alternative, evaluation can also be performed by online experiments. In an online experiment, recommendations are presented to a real user and the evaluation is made upon what items the user consumed or liked. In this sense, recommenders can be effectively evaluated regarding the consumption and preferences of the users. However, training new recommenders or researching novel recommending techniques are costlier using online evaluation than offline experiments, because experiments involve real users.

More recently, there has been efforts to reduce experiments with users, incorporating models that simulate users' participation in research experiments. For instance, using a collection of user browsing data, Zhao et al. [39] developed a model for incorporating eye tracking information with RS without requiring eye tracking technology. The authors developed a new click model for a ranking algorithm, and used as a signal of user satisfaction in Information Retrieval evaluation.

Moreover, there has also been a research in estimating the performance of costly controlled trials. Rosenfeld et al. [30] proposed a framework for predicting counterfactuals using a large historical data and small set of randomized trials. For instance, the authors predicted the Click-Through-Rate of a Search Engine by using a dataset of A/B tests.

The eye tracking and counterfactuals predictions for evaluation has been investigated in few Information Retrieval scenarios. Both researches were not sufficiently observed in Recommender System area. In both cases, online evaluation could be estimated, reducing costs of performing experiments with real users. In this sense, such strategies could be helpful in new algorithm designs focusing on evaluation of user satisfaction and user decision making process.

There are few work in the literature comparing the offline metrics with online evaluation. In Maksai et al. [23], the authors use offline metrics to predict online performance, using concepts such as coverage, diversity and serendipity. Besides this work, currently we see little correlation between the offline metrics and the online evaluation. CTR is a major online evaluation metric and sometimes, the utility based metrics are also used in online evaluation. Besides utility, it is unknown whether users' notions of novelty, unexpectedness and serendipity are the same to the evaluations performed by the current offline metrics. It is necessary a study to establish the level of consistency that the reviewed offline

metrics have with the online scenarios. Such study could perform online experiments with users in order to retrieve their feedback with regard to their perceived novelty, unexpectedness and serendipity towards personalized predictions. In this sense, the users' perception of these concepts can be correlated to the values obtained by the current metrics. Thus, the most correlated metrics can be used for both offline and online experiments, obtaining the closest values of novelty, unexpectedness and serendipity. Moreover, the users' feedback can also be used to design metrics for the life-level of novelty, for which there is currently no metric.

## 5 Conclusion

This paper performed a survey on the concepts of evaluation in RSs. As main contribution of this work, we surveyed 25 major state-of-art metrics and summarized them in six concepts: *utility, novelty, diversity, unexpectedness, serendipity* and *unexpectedness*. For each one, we presented the main definitions and existing metrics to evaluate the recommendations. Moreover, we also contributed by unifying the metrics under same notation and discussed the differences in definitions and metrics for each concept. For example, we classified novelty in three levels according to different definitions and we showed distinct and possibly uncorrelated metrics for serendipity and unexpectedness.

As another important contribution of this work, we looked inside the definitions and metrics of the concepts and we categorized them according to their user dependency characteristics. We also investigated the relationships among the concepts. However, in a further discussion, we showed that there might be unexplored correlations between the concepts that is worthy to study.

Our last contribution is a discussion of future research on evaluations in Recommendation. We point out three potential research directions: (1) re-examine the intentions of the concepts and their relationships; (2) factors that impact user satisfaction with recommendation; (3) methodology of online evaluation and closer correlation between offline metrics and online ones. These three directions are important and valuable for future work.

As potential future work, in Sect. 2.7 we briefly mentioned about Group Recommendation from one of our sources. We are not diving into this topic, however, it could be a possible future work on studying concepts and metrics of user satisfaction for Group Recommendation.

To highlight a final future work, an analysis on the evaluation of user satisfaction concepts and metrics for real practical recommender systems. Usually, the metrics are proposed by academic researchers or in co-operation with the industry. The aforementioned concepts in metrics are more common to the used in the literature. Currently, major metrics that are used in the industry are utility based ones, such as CTR, CTR-variance, retention and accuracy. Besides utility, coverage has receiving more attention in industry rather than in the academy. But they are also aware of the importance of the other mentioned concepts, for instance Gomez-Uribe and Hunt [13] mention the use of diversity and relevance. They are communicating with the academy on how to leverage the metrics in their systems. It will be an interesting issue to see how the user satisfaction metrics and others help the industry in the future.

**Acknowledgements** This work was supported by Natural Science Foundation (61532011, 61672311) of China and National Key Basic Research Program (2015CB358700).



## Appendix

Summarization of the metrics, categorization and their advantages and issues that have been mentioned in this paper (Tables 3, 4).

**Table 3** Summarization of the metrics

| Equations | Metric | Category | Advantages | Issues |
|---|---|---|---|---|
| 1 | $util(R_u) = MAE = \frac{\sum_{i \in R_u} p(i) - r(i)}{\lvert R_u \rvert}$ | Utility | Popular and simple<br>System's accuracy | |
| 2 | $util(R_u) = RMSE = \sqrt{\frac{\sum_{i \in R_u} (p(i) - r(i))^2}{\lvert R_u \rvert}}$ | Utility | Popular and simple<br>System's accuracy | Penalize large errors |
| 3 | $util(R_u) = precision = \frac{\lvert C_u \cap R_u \rvert}{\lvert R_u \rvert}$ | Utility | Popular and simple | |
| 4 | $util(R_u) = recall = \frac{\lvert C_u \cap R_u \rvert}{\lvert C_u \rvert}$ | Utility | Popular and simple | |
| | ROC curves | Utility | Easy to visualize | |
| 5 | $util(R_u) = rank(R_u) = \sum_{j=1}^{\lvert R_u \rvert} \frac{\max(r(i_j) - d, 0)}{2^{\frac{j-1}{\alpha - 1}}}$ | Utility | Useful when considering ranking and browsing order | Not so straightforward for many applications |
| 6 | $util(R_u) = CTR = \frac{\lvert C_u \rvert}{R_u}$ | Utility | Popular and Simple<br>Business metric | Online experiments are costly |
| 7 | $util(R_u) = \Delta_{retention} = p_t - p_c$ | Utility | Business metric | Online experiments are costly |
| 8 | $nov(R_u) = \sum_{i \in R_u} min_{j \in H_u} d(class(i), class(j))$ | Novelty level 2 | Simple and straightforward | Classes of items may not be available for all recommendation scenarios |
| 9 | $nov(R_u) = \sum_{i \in R_u} \frac{log_2 pop(i)}{\lvert R_u \rvert}$ | Novelty level 2 | Simple<br>Largely used | |
| 10 | $nov(R_u) = 1 - \frac{\lvert pop(i) \rvert}{\lvert U \rvert}$ | Novelty level 2 | Simple<br>Largely used | |
| 11 | $nov(R_u) = \frac{1}{\lvert R_u \rvert - 1} \sum_{j \in R_u} 1 - d(i,j)$ | Novelty level 3 | | More used for intra-list similarity. Similar to diversity metric |
| 12 | $nov(R_u) = \sum_{k=1}^{\lvert R_u \rvert} disc(k)(1 - p(seen \lvert i_k))$ | Novelty level 2–3 | Takes the rank into consideration | Not so straightforward for many applications |
| 13 | $div(R_u) = \sum_{i \in R_u} \sum_{j \in R_u, i \neq j} d(i,j)$ | Diversity | Simple and straightforward<br>Largely used | |
| 14 | $div(R_u) = \sum_{i \in R_u} \sum_{j \in R_u, i \neq j} cossim(i,j)$ | Diversity | Simple and straightforward<br>Largely used | |
| 15 | $div(R_u) = \sum_{k=1}^{\lvert R_u \rvert} \sum_{l=1}^{\lvert R_u \rvert} disc(k) disc(l \lvert k) d(i_k, i_l) \ \forall i_k \neq i_l$ | Diversity | Takes the rank into consideration | Not so straightforward for many applications |
| 16 | $unexp(R_u) = R_u - PM_u$ | Unexpectedness | Measures unexpectedness as a deviation from expected recommendations | Depend on defining a set of expected recommendations for the scenario |
| 17 | $unexp(R_u) = \frac{R_u - PM_u}{\lvert R_u \rvert}$ | Unexpectedness | Measures unexpectedness as a deviation from expected recommendations | Depend on defining a set of expected recommendations for the scenario |
| 18 | $unexp(R_u) = \sum_{i \in R_u} \sum_{j \in H_u} PMI(i,j)$ | Unexpectedness | Calculates the probability of items to be rated by the user | It can be biased towards rare items |
| 19 | $unexp(R_u) = \sum_{i \in R_u} max_{j \in R_u} PMI(i,j)$ | Unexpectedness | Calculates the probability of items to be rated by the user | It can be biased towards rare items |
| 20 | $unexp(R_u) = \frac{1}{\frac{1}{\lvert F_i \rvert} \sum_{v,w \in F_i} \frac{I_v}{I_V + I_w - I_{v,w}}}$ | Unexpectedness | | Unpersonalized metric |
| 21 | $ser(R_u) = \sum_{k=1}^{\lvert R_u \rvert} \max(R_u[k] - PM_u[k], 0) rel(i_k) \frac{count_k(k)}{k}$ | Serendipity | Uses utility and surprise<br>Considers rank | Metric depends on defining a set of expected recommendations for the scenario |
| 22 | $ser(R_u) = \frac{\sum_{i \in UNEXP_u} \text{utility(i)}}{\lvert R_u \rvert}$ | Serendipity | Uses utility and surprise | Metric depends on defining a set of expected recommendations for the scenario |
| 23 | $ser(R_u) = \frac{(R_u - E_u) \cap USEFUL_u}{\lvert R_u \rvert}$ | Serendipity | Uses utility and surprise | Metric depends on defining a set of expected recommendations for the scenario |

**Table 3** (continued)

| Equations | Metric | Category | Advantages | Issues |
|---|---|---|---|---|
| 24 | $ser(R_u) = \frac{1}{\lvert H_u \rvert} \sum_{i \epsilon H_u} \sum_{j \epsilon R_u} \frac{cossim(i,j)}{\lvert R_u \rvert}$ | Serendipity | Uses novelty and surprise | It is more related to a novelty metric<br>Metric does not consider utility |
| 25 | $cov = \frac{\lvert I_p \rvert}{I}$ | Coverage | Metric captures the proportion of items in the system which is able to be predicted | |
| 26 | $cov = \frac{\lvert \sum_{j=1..n} I_L^j \rvert}{\lvert I \rvert}$ | Coverage | Metric captures the proportion of items in the system which is able to be predicted | |
| 27 | $cov = \frac{1}{\lvert I \rvert - 1} \sum_{j=1}^{\lvert I \rvert} (2j - \lvert I \rvert - 1) p(I_j)$ | Coverage | Captures coverage and the disparity of recommendation among different items | It is a more complex metric |

**Table 4** Summarization of the performances of state-of-art Recommender Systems in terms of different metrics

| Reference | Metric (equation) | Evaluation methodology | Algorithm | Database | Result | |
|---|---|---|---|---|---|---|
| Wen [37] | RMSE (2) | Train-test split | Item KNN | Netflix | 0.95 | |
| | | | Item EM | | 0.91 | |
| | | | SVD | | 0.8974 | |
| Vargas and Castells [36] | Novelty level 3 (10) | Fivefold CV | MF | Movie Lens | 0.058 | |
| | | | IA-select adaptation | | 0.0639 | |
| | | | MMF | | 0.0620 | |
| | | Temporal train-test split | MF | Last FM | 0.2671 | |
| | | | IA-select adaptation | | 0.3462 | |
| | | | MMF | | 0.2439 | |
| Vargas and Castells [36] | Diversity (13) | Fivefold CV | MF | Movie lens | 0.0471 | |
| | | | IA-select adaptation | | 0.0537 | |
| | | | MMF | | 0.0510 | |
| | | Temporal train-test split | MF | Last FM | 0.2518 | |
| | | | IA-select adaptation | | 0.3343 | |
| | | | MMF | | 0.2360 | |
| De Pessemier et al. [8] | Diversity (11) | Train-test group recommendation | User-KNN | Movie lens | 0.7 | |
| | | | Item-KNN | | 0.81 | |
| | | | SVD | | 0.64 | |
| Adamopolous et al. [1] | Unexpectedness (15) | Train-test | Item KNN, User KNN, MF | Movie lens | 0.71 (Top10) | 0.75 (Top100) |
| | | | | Book Crossing | 0.3 (Top10) | 0.38 (Top100) |
| Lu et al. [22] | Serendipity (21) | Train-test | SVD | Netflix | 0.2534 | |
| | | | SVD++ | | 0.3036 | |
| | | | SVDNbr | | 0.2978 | |
| | | | SVD | Yahoo! Music | 0.1995 | |
| | | | SVD++ | | 0.2771 | |
| | | | SVDNbr | | 0.3834 | |
| Shi et al. [33] | RMSE (2) | Temporal train-test steps split | Item KNN | Movie lens | 1.05 (time step 1) | 1.65 (time step 40) |
| | Novelty (9) | | | | ~10,000 (time step 1) | ~18,000 (time step 40) |
| | Diversity (11) | | | | ~0.025 (time step 1) | ~0.035 (time step 40) |
| Adamopolous et al. [1] | Catalog coverage (24) | Train-test | ItemKNN, UserKNN, MF | Movie Lens | 0.05 (top 10) | 0.15 (top 100) |
| | | | | Book Crossing | 0.2 (top 10) | 0.5 (top 100) |

## References


1. Adamopoulos P, Tuzhilin A (2014) On unexpectedness in recommender systems: or how to better expect the unexpected. ACM Trans Intell Syst Technol 5(4):Article 54
2. Adomavicius G, Tuzhilin A (2005) Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Trans on Knowl Data Eng 17(6):734–749
3. Akiyama T, Obara K, Tanizaki M (2010) Proposal and evaluation of serendipitous recommendation method using general unexpectedness. In: Workshop on the practical use of recommender systems, algorithms and technologies
4. Bobadilla J, Ortega F, Hernando A, Gutiérrez A (2013) Recommender systems survey. Knowl Based Syst 46:109–132
5. Celma O, Herrera P (2008) A new approach to evaluating novel recommendations. In: Proceedings of the 2008 ACM conference on recommender systems (RecSys '08). ACM, New York, NY, USA, pp 179–186
6. Celma O (2010) The long tail in recommender systems. Springer, Berlin, pp 87–107
7. Chu W, Park S (2009) Personalized recommendation on dynamic content using predictive bilinear models. In: Proceedings of the 18th international conference on World wide web (WWW '09). ACM, New York, NY, USA, pp 691–700
8. De Pessemier T, Dooms S, Martens L (2014) Comparison of group recommendation algorithms. Multimed Tools Appl 72(3):2497–2541
9. Dror G, Koenigstein N, Koren Y, Weimer M (2011) The Yahoo! music dataset and KDD-Cup'11. In: Dror G, Koren Y, Weimer M (eds) Proceedings of the 2011 international conference on KDD Cup 2011-volume 18 (KDDCUP'11), vol 18. JMLR.org, pp 3–18
10. Farris PW, Bendle NT, Pfeifer PE, Reibstein D (2010) Marketing metrics: the definitive guide to measuring marketing performance. Pearson Education, Inc., Upper Saddle River **(ISBN 0–13-705829-2)**
11. Gemmis M, Lops P, Semeraro G, Musto C (2015) An investigation on the serendipity problem in recommender systems. Inf Process Manag 51(5):695–717
12. Ge M, Delgado-Battenfeld C, Jannach D (2010) Beyond accuracy: evaluating recommender systems by coverage and serendipity. In: Proceedings of the fourth ACM conference on recommender systems (RecSys '10). ACM, New York, NY, USA, pp 257–260
13. Gomez-Uribe C, Hunt N (2015) The Netflix recommender system: algorithms, business value, and innovation. ACM Trans Manag Inf Syst 6(4):Article 13
14. Herlocker J, Konstan J, Terveen L, Riedl J (2004) Evaluating collaborative filtering recommender systems. ACM Trans Inf Syst 22(1):5–53
15. Hurley N, Zhang M (2011) Novelty and diversity in top-n recommendation—analysis and evaluation. ACM Trans Internet Technol 10(4):Article 14
16. Iaquinta L, Gemmis M, Lops P, Semeraro G. Molino P (2010) Can a recommender system induce serendipitous encounters? In: Kang K (ed) E-commerce. InTech. https://www.intechopen.com/books/e-commerce/can-a-recommender-system-induce-serendipitous-encounters-
17. Iaquinta L, Gemmis M, Lops P, Semeraro G, Filannino M, Molino P (2008) Introducing serendipity in a content-based recommender system. In: Proceedings of the 2008 8th international conference on hybrid intelligent systems (HIS '08). IEEE Computer Society, Washington, DC, USA, pp 168–173
18. Kaminskas M, Bridge D (2014) Measuring surprise in recommender systems. In: Workshop on recommender systems evaluation: dimensions and design (REDD 2014), October 10, 2014, Silicon Valley, USA
19. Kapoor K, Kumar V, Terveen L et al (2015) "I Like to Explore Sometimes": adapting to dynamic user novelty preferences. In: Proceedings of the 9th ACM conference on recommender systems (RecSys '15). ACM, New York, NY, USA, pp 19–26
20. Kotkov D, Veijalainen J, Wang S (2016) Challenges of serendipity in recommender systems. In: WEBIST 2016: proceedings of the 12th international conference on web information systems and technologies, vol 2, pp 251–256
21. Kotkov D, Wang S, Veijalainen J (2016) A survey of serendipity in recommender systems. Knowl Based Syst 111:180–192
22. Lu Q, Chen T, Zhang W, Yang D, Yu Y (2012) Serendipitous personalized ranking for top-n recommendation. In: Proceedings of the 2012 IEEE/WIC/ACM international joint conferences on web intelligence and intelligent agent technology-volume 01 (WI-IAT '12), vol 1. IEEE Computer Society, Washington, DC, USA, pp 258–265
23. Maksai A, Garcin F, Faltings B (2015) Predicting online performance of news recommender systems through richer evaluation metrics. In: Proceedings of the 9th ACM conference on recommender systems (RecSys '15). ACM, New York, NY, USA, pp 179–186
24. McNee S, Riedl J, Konstan J (2006) Being accurate is not enough: how accuracy metrics have hurt recommender systems. In: CHI '06 extended abstracts on human factors in computing systems (CHI EA '06). ACM, New York, pp 1097–1101
25. Murakami T, Mori K, Orihara R (2008) Metrics for evaluating the serendipity of recommendation lists. In: Proceedings of the 2007 conference on new frontiers in artificial intelligence (JSAI'07). Springer, Berlin, Heidelberg, pp 40–46
26. Nakatsuji M, Fujiwara Y, Tanaka A et al (2010) Classical music for rock fans? Novel recommendations for expanding user interests. In: Proceedings of the 19th ACM international conference on Information and knowledge management (CIKM '10). ACM, New York, NY, USA, pp 949–958
27. Onuma K, Tong H, Faloutsos C (2009) TANGENT: A novel, 'Surprise Me', recommendation algorithm. In: KDD'09. ACM, New York, NY, USA, pp 657–666
28. Ribeiro M, Ziviani N, De Moura E et al (2014) Multi objective pareto-efficient approaches for recommender systems. ACM Trans Intell Syst Technol 5(4):Article 53
29. Ricci F, Rokach L, Shapira B, Kantor P (2011) Recommender systems handbook. Springer, Berlin
30. Rosenfeld N, Mansour Y, Yom-Tov E (2016) Predicting Counterfactuals from large historical data and small randomized trials. arXiv preprint arXiv:1610.07667
31. Schwartz B (2004) The paradox of choice: why less is more. Ecco, New York
32. Shani G, Gunawardana A (2009) Evaluating recommender systems. Microsoft Research. https://www.microsoft.com/en-us/research/publication/evaluating-recommender-systems/
33. Shi XY, Luo X, Shang MS, Gu L (2017) Long-term performance of collaborative filtering based recommenders in temporally evolving systems. In: Neurocomputing, vol 267, pp 635–643
34. Silveira T, Rocha L, Mourão F, Gonçalves M (2017) A framework for unexpectedness evaluation in recommendation. In: Proceedings of the symposium on applied computing (SAC '17). ACM, New York, NY, USA, pp 1662–1667
35. Vargas S, Baltrunas L, Karatzoglou A, Castells P (2014) Coverage, redundancy and size-awareness in genre diversity for recommender systems. In: RecSys'14. ACM, New York, pp 209–216
36. Vargas S, Castells P (2011) Rank and relevance in novelty and diversity metrics for recommender systems. In: RecSys'11. ACM, New York, NY, USA, pp 109–116
37. Wen Z (2008) Recommendation system based on collaborative filtering. http://www.zheng-wen.com/WenRecommendation.pdf

38. Zhang Y, Séaghdha D, Quercia D, Jambor T (2012) Auralist: introducing serendipity into music recommendation. In: WSDM'12. ACM, New York, NY, USA, pp 13–22
39. Zhao Q, Chang S, Harper F, Konstan J (2016) Gaze prediction for recommender systems. In: RecSys'16. ACM, New York, pp 131–138
40. Zhou T, Kuscsik Z, Liu J et al (2010) Solving the apparent diversity-accuracy dilemma of recommender systems. Proc Natl Acad Sci 107(10):4511–4515
41. Ziegler C, McNee S, Konstan J, Lausen G (2005) Improving recommendation lists through topic diversification. In: WWW'05. ACM, New York, NY, USA, pp 22–32