Uncovering Recommendation Serendipity with Objective Data-driven Factor Investigation

WENJUN JIANG, SONG LI, XUEQI LI, and KENLI LI, Hunan University, China
JIE WU, Temple University, USA

The serendipity recommendation tries to burst the filter bubble while still meeting user interests. However, serendipity itself has not been well understood in the recommendation system. Thus, factor investigation in recommendation serendipity has attracted much attention, for which two challenges hinder follow-up research: (1) *Ambiguity of factors*. Different works exploit different factors, and the meanings of factors are inconsistent in various works. (2) *Lack of complete impact validation*. The importance of these factors in different domains is not yet fully understood. The common approach of user surveys costs much, but the results are usually less objective and limited in quantity. To this end, we strive to comprehensively identify and clarify serendipity factors and explore objective data-driven approaches to validate factor impacts in large-scale cross-domain scenarios. We first conduct a comprehensive literature review to identify all possible factors, from which we find that some factors are being used indistinguishably. To address this issue, we propose two principles of meaning coverage and factor independence to clarify and disentangle serendipity factors. Next, we propose a general experimental framework to explore the impacts of factors. Then, we implement one such framework and run experiments on nine representative datasets to study factor importance on serendipity. We also propose a quantitative method to measure the degree of disentanglement of factors and to test the effects of factor combinations. We gain several useful findings: (1) *relevance*, *diversity*, and *random* are critical factors affecting serendipity; (2) domain features affect factor importance and can guide serendipity recommendation; (3) the disentanglement quantification method benefits the understanding of serendipity and the combination of factors. To our knowledge, this is the first work to comprehensively investigate serendipity factors and experimentally compare their impacts in an objective data-driven approach.

CCS Concepts: • **Information systems → Social recommendation**; **Personalization**.

Additional Key Words and Phrases: Serendipity Recommendation, Factor investigation, Objective Data-driven, Experimental Framework

ACM Reference Format:
Wenjun Jiang, Song Li, Xueqi Li, Kenli Li, and Jie Wu. 2025. Uncovering Recommendation Serendipity with Objective Data-driven Factor Investigation. 1, 1 (July 2025), 31 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

As an important web mining technique for alleviating information overload, recommendation systems help users find the right items from many candidates efficiently. Several advanced techniques have been proposed to improve the accuracy of the recommendation, e.g. [12, 15, 16, 18, 20, 80, 82, 95, 97, 100, 105, 106]. However, accuracy-oriented recommendation aggravates the information cocoons in which users can only see items that match their profiles or past

This research was supported by the National Natural Science Foundation of China (Grant No. 62172149).
Authors' addresses: Wenjun Jiang, jiangwenjun@hnu.edu.cn; Song Li, leesong@hnu.edu.cn; Xueqi Li, lee_xq@hnu.edu.cn; Kenli Li, lkl@hnu.edu.cn, Hunan University, College of Computer Science and Electronic Engineering, 116 Lu Shan South Road, Changsha, Hunan, China, 410082; Jie Wu, jiewu@temple.edu, Temple University, Department of Computer and Information Sciences, Philadelphia, PA, 19122, USA.

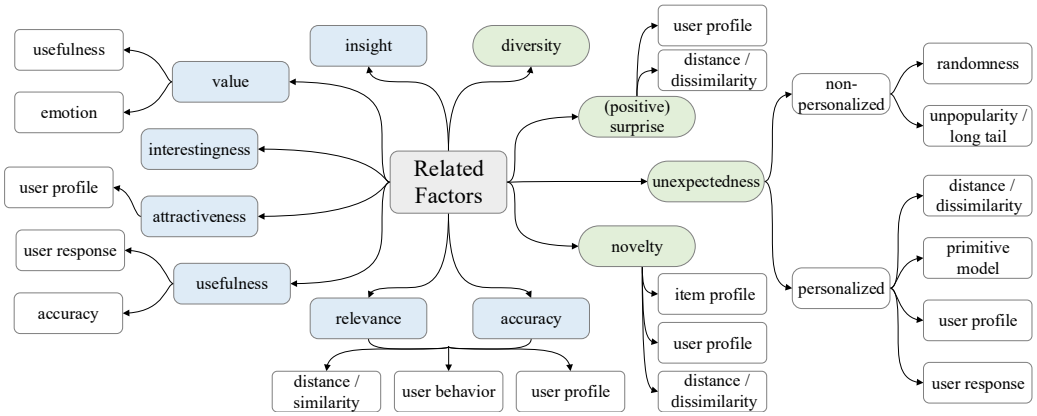


Fig. 1. Overview of Frequently Mentioned Serendipity Related Factors in Literature. The Right (Green Box) for Bursting Filter Bubbles; The Left (Blue Box) for Matching User Interests.

interests [27, 88, 89], which narrows users' horizons, exacerbates the filter bubbles [25], and aggravates the long-tail effect of platforms [21]. "Serendipity is an increasingly recognized design principle of the infosphere, and pursuit for serendipity can help to burst filter bubbles and weaken echo chambers [75]." Therefore, the serendipity recommendation is introduced [8, 22, 35, 65, 108], which tries to find items that interest users, but are beyond their discovery [23, 51, 52, 70]. However, the subjectivity nature makes it difficult to directly define and evaluate the serendipity of the recommendation [61, 108]. Thus, researchers turn to study various factors of serendipity in recommendation systems.

We summarize the factors most frequently mentioned as shown in Fig. 1 and Table 5. It shows that various factors have been considered in different works, making it difficult to reach a consensus for recommendation serendipity. Furthermore, the meanings and names of serendipity factors are inconsistently used, i.e., some factors with the same name have different meanings and implementations, while some factors share the same meanings but are given different names. This further exacerbates the chaos of conception. Taking *unexpectedness* in Fig. 1 as an example, in literature it has two non-personalized and four personalized ways/bases to implement, and some of them overlap with *novelty* and *surprise*. If some platform or research would like to optimize *unexpectedness*, it must determine a definite meaning of this factor; this also applies to all other factors. Going a further step, if it comes to optimizing all possible serendipity factors, a set of non-overlapped factors that cover all meanings will be necessary for the efficiency and effectiveness, to avoid redundancy or omission. This is consistent with the MECE rule (Mutually Exclusive, Collectively Exhaustive) [64], an effective framework for categorizing ideas/tasks/issues to cover all options without overlapping. Therefore, it is urgent to clarify and disentangle serendipity factors to help understand recommendation serendipity, benefit downstream tasks, and promote future studies.

Moreover, it is also challenging to study the impact of various factors on recommendation serendipity. Conducting user surveys is an effective validation method. Kotkov et al. [42] surveyed 475 users about eight descriptions of serendipity with 2,146 movies on MoVielens[1]. Chen et al. [13] conduct a user survey involving 3,039 users on Taobao[2] to study serendipity to improve user satisfaction. Wang et al. [87] study what item features and user characteristics might affect perceived serendipity of users with a large-scale user survey, involving more than 10,000 users on mobile

[1]https://movielens.org
[2]https://www.taobao.com/

Taobao. Most recently, Denis et al. [43] collected 2002 survey responses from 397 users of an online article recommender system and studied the impact of commonly mentioned factors on three types of serendipity: RecSys serendipity, user serendipity, and generalized serendipity. Survey-based findings help to understand what is related to the serendipity of recommendations. However, user surveys usually cost a lot of time and energy and are limited in users, items, domains, and factors. Moreover, users may become impatient and their responses may be arbitrary or involve judgment bias [1].

In summary, there are two challenges in factor investigation for recommendation serendipity: (1) *Ambiguity of factors*. Different works exploit different factors, and the meanings of factors are inconsistent in various works, which hinders follow-up research. Therefore, it is urgent to comprehensively identify and clarify serendipity factors. (2) *Lack of complete impact validation*. The roles of factors in the recommendation serendipity of various domains are still not fully understood. User surveys cost a lot, while results are usually less objective and limited in quantity. Therefore, it is necessary to try other approaches that involve less human intervention to validate factor impacts in large-scale cross-domain scenarios.

Our Motivations. Keeping in mind the goal of understanding the serendipity of recommendation and the above challenges, our motivations are threefold. (1) Exploring what factors are taken as related to the serendipity of recommendations in literature and clarifying the meanings of serendipity factors. In particular, the name and meaning of each factor should be consistent and unique, and there should be less overlap between any two factors to avoid ambiguity and to benefit future studies. (2) Validating the impact of each factor on the serendipity of the recommendation in an objective data-driven approach. The validation method should involve less human intervention and can be flexibly applied in various domains. (3) Quantitative measurement of the degree of disentanglement between factors and exploration of the proper ways to effectively integrate different factors to improve serendipity recommendation.

Our contributions. We strive to explore an objective approach with large-scale cross-domain and publicly available datasets where less human intervention is involved, to uncover what factors contribute more to recommendation serendipity and validate factor impacts in different domains. Our contributions are threefold.

(1) **Identifying and clarifying factors of recommendation serendipity**. We comprehensively review the literature to explore all possible factors. To address ambiguity issues, we propose two principles of meaning coverage and factor independence to clarify and disentangle factors. We gain a set of seven factors for recommendation serendipity, including three personalized factors: *relevance*, *difference*, and *diversity*, and four non-personalized factors: *novelty*, *unpopularity*, *high quality*, and *random*. This work will benefit many downstream tasks, e.g., facilitating interactive recommendations by simplifying users' factor selection and enhancing the explainability by providing the scores of the recommended items on some key factors. (Section 3)

(2) **Framework for evaluating the impact of factors on the recommendation serendipity**. We propose a general experimental framework and provide a specific implementation to validate the impact factor on recommendation serendipity. It has two components: the optimization strategy for each factor and the factor impact evaluation with comparative performance. It can be applied to check the importance of all the factors that users and service providers may be concerned about. We also introduce an innovative quantitative method to assess the degree of disentanglement between factors. It helps quantify information redundancy and select the most important factors for serendipity in different domains, providing a novel perspective and theoretical support to optimize serendipity recommendations. (Section 4)

(3) **Extensive experiments of factor impacts on recommendation serendipity**. The experimental results on nine representative datasets validate that the proposed optimization strategies are effective and reflect impact factors in different domains. In general, *relevance*, *diversity*, and *random* have more impacts on serendipity. Meanwhile, the role of

factors varies with the datasets. We analyze the reason and find that the domain features also matter. This can help guide the serendipity recommendation. For instance, if users' interest is more focused in some domain, then *relevance* can be given more weight; and vice versa. In addition, we also check the correlation among factors and test the effects of factor combinations based on the degree of disentanglement. (Section 5)

To our knowledge, this is the first work that comprehensively investigates and disentangles factors of recommendation serendipity and experimentally analyzes their impacts in an objective data-driven approach. The proposed factor investigation method and experimental framework are general and can deal with new factors. The remainder is organized as follows. Section 2 provides some representative research on serendipity recommendation in different domains. The details of the factor investigation are described in Section 3. Section 4 introduces the experimental framework for evaluating factor impacts, implementing such a framework, and defining and calculating the degree of disentanglement between factors. Extensive experiments with nine real-life datasets are presented and analyzed in Section 5. Finally, Section 6 concludes this paper and highlights future directions.

2 SERENDIPITY RECOMMENDATIONS IN DIFFERENT DOMAINS

We briefly review the literature on serendipity recommendations, particularly those methods in different domains.

Much research has been conducted on e-commerce. Wang et al. [86] proposed the INVBCF algorithm to recommend serendipitous cold items to innovators in the e-commerce environment. Zhang et al. [98] studied the effect of price on predicting the probability of user purchase for a session-based recommendation in e-commerce. Wang et al. [90] proposed an industrial framework for serendipity recommendation in e-commerce, in which unexpected and satisfying items are selected to achieve both accuracy and novelty. Some works are particularly done for movie recommendations. Yang et al. [93] identified serendipitous items with post-satisfaction and pre-interest for movie recommendation. Leung et al. [49] built the emotion-aware recommendation system to help with serendipity recommendation for movies. Nugroho et al. [68] studied user curiosity in determining recommendation serendipity and also for movie recommendation, which considered relevance, novelty, popularity, unexpectedness, and diversity.

Some works have been done for serendipity recommendations on scientific collaboration, articles, courses, tourism, or trips. Wan et al. [85] took unexpectedness, value, and relevance into account to identify serendipitous collaborators for scientific collaboration. Magara et al. [58] exploited latent relationships to recommend serendipitous research articles. Pardos and Jiang [70] designed a serendipity model for a university course recommendation system. Menk et al. [62] proposed an online tourism recommendation system by exploiting the curiosity of users of online social networks. Gu et al. [32] proposed a trip recommendation model considering POIs and attractive routes, which identified the attractive routes with the popularity and Gini coefficient of POIs.

All of the above representative works indicate that serendipity and its related factors can benefit personalized recommendations in many domains. Meanwhile, they validate the existence of different factors being considered for the serendipity of recommendation and the inconsistency between factor names and meanings, which motivates our work in this paper.

3 FACTOR INVESTIGATION

According to the systematic review method of the literature [89, 108], we have thoroughly researched the literature to extract factors related to serendipity. Then we summarized frequently mentioned factors or components of serendipity in the recommendations. We found that in various works, factors with the same name may have different meanings; meanwhile, the same meaning might be expressed by different factor names. The inconsistency between the factor

names and their meanings hinders follow-up explorations. Incited by the effective framework of the MECE rule [64] to categorize ideas/tasks/issues to cover all options without overlapping, we propose two principles to clarify and disentangle those serendipity factors: (1) covering the meanings of all frequent factors in the literature, and (2) keeping factors independent of each other. The processes of factor extraction, analysis, and clarification are described below.

3.1 Factor Extraction

Based on keyword search and text matching, we selected articles from DBLP[3], Google Scholar[4], as well as IEEE and ACM digital libraries on serendipity and serendipity recommendation. To be specific, on DBLP, we collect all of the papers using the keywords "serendipity", "serendipi recommend", "surpris recommend", "novel", "unexpect", where "serendipi" denotes "serendipity" and "serendipitous", "recommend" for "recommend", "recommendation" and "recommender", "surpris" for "surprise", "surprising" and "surprised", and "unexpect" for "unexpectedness" and "unexpective"; On Google Scholar, we select the first 100 papers with the query "(serendipity OR serendipitous OR novel OR unexpect) AND (recommend OR recommender OR recommendation)" and select the first 50 papers with the query "(surprise OR surprising OR surprised) AND (recommend OR recommender OR recommendation)". In IEEE and ACM digital libraries, we use similar keywords and choose papers on the topic of information retrieval and recommender systems.

We take the related works since 2016 as the main references. We also include some earlier representative works, e.g., [3, 26, 35, 39, 56, 60, 65, 69, 101]. Moreover, we refer to several survey papers (e.g. [1, 2, 22, 30, 38, 45, 75, 79, 108]) to ensure that more possible factors are extracted. Finally, we collected more than 130 articles that use serendipity factors as main references. Moreover, our method can deal flexibly with new factors, in case some factors are missed or new factors are identified in the future.

Table 5 shows some commonly mentioned serendipity factors in the literature. According to the literature review, most researchers believe that *unexpectedness* and *relevance* contribute greatly to the serendipity of recommendation [7, 13, 17, 27, 42, 44, 45, 57, 66, 70, 71, 76, 81, 96, 107]. Some works evaluate it with user feedback [4, 13, 42]. In [42, 44, 71], *novelty* is also a critical factor. In [13], *timeliness* is taken as a more important factor than *unexpectedness*. Moreover, *value* [92], *positive surprise* [27, 84], *diversity* [17] and *temporal diversity* [48, 99] are also being taken as important factors of serendipity. Another factor combination consists of *unexpectedness* and *usefulness* [5, 26, 30, 59], and *insight* is also mentioned [104].

To extract the main factors from various combinations, we count the frequency of the factors mentioned. The most frequently mentioned two factors are *unexpectedness* ($\geq$ 30 times) and *relevance* ($\geq$ 15 times). Then follow *surprise* ($\geq$ 8), *novelty* ($\geq$ 8), *usefulness* ($\geq$ 7), *value* ($\geq$ 7), etc. According to the goals of factors, we divide the commonly used factors into two categories, one for bursting filter bubbles and the other for matching user interests (ability). The former consists of four factors, *unexpectedness*, *surprise*, *novelty*, and *diversity*; and the latter consists of seven factors, *accuracy*, *relevance*, *usefulness*, *attractiveness*, *interestingness*, *value*, and *insight*. We also summarize those factors and their meanings (implementations or bases) in boxes and display them clockwise as shown in Fig. 1.

3.2 Factor Analysis

We analyze serendipity factors from two categories: factors for bursting filter bubbles and for matching user interests.

[3]https://dblp.org
[4]https://scholar.google.com

3.2.1 Factors for Bursting Filter Bubbles. Many serendipity factors are proposed for bursting filter bubbles. The four most important are *unexpectedness*, *surprise*, *novelty*, and *diversity*.

As shown in Fig. 1, the meanings of **unexpectedness** vary in different works, which could be divided into non-personalized and personalized concepts, as follows:

- In non-personalized ways, Ge et al. [27] proposed that randomness and non-determinism algorithms could achieve *unexpectedness*. Silva et al. [81] regarded items with fewer ratings as unexpected ones. Maksai et al. [59] penalized *expectedness* of the most popular items, and [30, 107] thought items in the long tail would be unexpected.
- In personalized ways, four bases are used for *unexpectedness*: distance/similarity between the user and item, the primitive model, the user profile, and the response of the user. Some works utilized distance or similarity to define *unexpectedness*, treating dissimilar items [17, 36, 41, 44, 67, 84, 96] or those with a certain distance range as unexpected [7, 19, 50, 57]. Some researchers defined *unexpectedness* based on a primitive model [26, 29, 56, 65, 66, 69, 73], where items not recommended by the primitive model are seen as unexpected ones. User profiles are also used to define *unexpectedness* [42, 45, 79], e.g., research topics [92] and user's latest preferences [104]. Moreover, some researchers thought that *unexpectedness* is determined by the user response [11, 76].

The factor of **surprise** expresses a similar meaning to *unexpectedness*, that is, a violation of the user expectation [67, 84] and far from the user profile [79].

For the factor of **novelty**, an item can be novel to users in three ways [39]: new to the system, new to the user, and forgotten items. [1, 13, 38, 44, 45] define *novelty* based on user profiles, where items different from user profiles are taken as novel ones. [74] regards an item to be novel when more users have never interacted with it; while they are regarded as unexpected in some other works.

Recently, **diversity** [17, 28, 103] has been employed to improve the difference among recommendations. [48] investigates *temporal diversity* which means the temporal characteristics of top-k recommendations, [47] analyzes how *diversity* affects user purchase preferences, [77] improves *diversity* with weak ties and [24] improves it with knowledge graphs. [54] captures feature-aware diversity with a feature-disentanglement self-balancing re-ranking framework.

3.2.2 Factors for Matching User Interest. Important factors for matching user interest include *value*, *usefulness*, *accuracy*, *relevance*, *attractiveness*, *interestingness* and *insight*.

The factor of **value** involves emotional aspects [9, 67] and *usefulness* [1, 36, 67, 104]. While **usefulness** could be determined by user response [26] or *accuracy* [59]. And **accuracy** is similar to *relevance* [41, 52, 96]. While **relevance** is employed to improve the similarity between recommendations and user profile [17, 76], historical items [45, 96, 107] (or those with higher ratings [44, 56, 81]) and target users in latent space [92]. In addition, the factors of **attractiveness** [29, 47] and **interestingness** [34, 40] denote similar meanings, i.e., the closeness to a user profile or history. The factor of **insight** denotes the ability to connect recommendations with users [73, 104].

Main issues. We find that in various works, the same factor name usually expresses different meanings and vice versa. As shown in Fig. 1, six types of *unexpectedness* and three types of *relevance* are proposed. The inconsistency between the names and meanings hinders the follow-up research. Hence, there is an urgent need to clarify these factors.

3.3 Factor Clarification

We represent each factor f_i as a two-tuple $(name_i, meaning_i)$. The goals of factor clarification are threefold, as follows:

Goal 1: The name and meaning of each factor should be one-to-one correspondence, i.e., $name_i \leftrightarrow meaning_i$;
Goal 2: The meanings of any two factors should have no overlap, i.e., if $i \neq j$, then $meaning_i \bigcap meaning_j = \emptyset$;
Goal 3: The sum of meanings and corresponding implications inherent in all factors should remain unchanged, i.e., $Sum_{meaning} = \sum_{i \in [1,n]} meaning_i^{original} = \sum_{j \in [1,m]} meaning_j^{after}$.

Based on the principles of meaning coverage and factor independence, we perform disentanglement of the 11 frequently mentioned factors displayed in Fig. 1 with the following three steps. Other factors that are not frequently mentioned or will be identified in the future can also be treated incrementally. Suppose $f_i = (name_i, meaning_i)$ is the original factor that needs to be processed. Initially, let the set of disentangled factors (i.e., factors obtained by disentanglement) as null and $Sum_{meaning}$ as empty. We conduct disentanlgement mainly according to the factor meanings, as follows:

Step 1: If its meaning is contained by existing disentangled factors, i.e., $meaning_i \subseteq Sum_{meaning}$, it would be discarded or merged directly;
Step 2: If there is some overlap between $meaning_i$ and $Sum_{meaning}$, then $meaning_i - (meaning_i \bigcap Sum_{meaning})$, the non-overlapping part, would be taken into $Sum_{meaning}$ and given a name different from other disentangled factors (can be $name_i$);
Step 3: Otherwise, if $meaning_i \bigcap Sum_{meaning} = \emptyset$, $factor_i$ can be taken as a new factor, i.e., $meaning_i$ would be taken into $Sum_{meaning}$ and given a name different from other disentangled factors (can be $name_i$).

Note that a new factor may also be treated in other ways, e.g., keeping its meaning and fragmenting or rearranging a previously adopted meaning/factor in light of new information. The specific approach can be considered on a case-by-case basis.

3.3.1 Factor Disentanglement. We deeply analyze factor relations, especially their overlap in meanings, then clarify and disentangle them. The process is somewhat similar to database normalization [46, 78], a systematic approach of decomposing tables to eliminate data redundancy and ensure that data dependencies make sense. We can take the factor list as a table with two columns: name and meaning. The goal of disentanglement is to make each name correspond to a meaning that does not overlap with other factors. Fig. 2 and Fig. 3 illustrate the disentanglement of factors for bursting filter bubbles and those for matching user interest, respectively. They show that the original factors have several implementations/meanings and overlap heavily (the left two columns); while after disentanglement, the factors are independent of each other, and all meanings are covered (the right two columns).

In the following, we try to disentangle factors with the principles of “meaning coverage" (i.e., make the disentangled factors contain all meanings) and “factor independence" (i.e., keep no overlap among the factors). We conduct disentanglement mainly with the original meanings in the second column in Fig. 2 and Fig. 3. Note that the factor meanings based on user response or emotion are taken as interactive factors, which will be left for future study.

We first deal with the four factors in Fig. 2. There are six types of *unexpectedness*: based on randomness, unpopularity, user response, primitive model, user profile, and user-item dissimilarity, respectively. Meanwhile, the factor of *novelty* involves three bases: user-item dissimilarity, user profile, and item profile. Moreover, *surprise* has two meanings based on the user-item dissimilarity and the difference from the user profile, respectively, which are also contained in *unexpectedness* and *novelty*. Finally, *diversity* has two meanings based on the user-item dissimilarity and the item profile, respectively.

We treat non-interactive meanings with our disentanglement principles. The third meaning, which is based on user response, is taken as interactive factors.

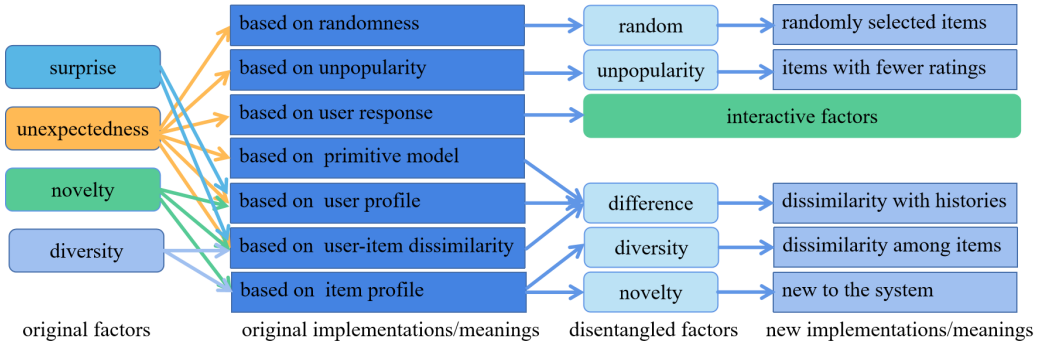


Fig. 2. Illustration of Factor Disentanglement: Factors for Bursting Filter Bubbles.

- For the first meaning, which is based on randomness, since this meaning has no overlap with disentangled factors (which is null currently), we explore Step 3 of the three steps of disentanglement, i.e., take all this meaning and take it as an independent factor, *random*, and assign it a more clear implementation/meaning as shown in the fourth column in Fig. 2.
- For the second meaning, which is based on unpopularity, since this meaning has no overlap with disentangled factors (i.e.,*random*), we explore Step 3 again, i.e., take all this meaning and take it as an independent factor, *unpopularity*, and assign it a more clear implementation/meaning.
- For the fourth meaning of *unexpectedness* based on the primitive model, since this meaning has no overlap with disentangled factors (i.e., *random* and *unpopularity*), we take it with a new factor of *difference*, and assign it a clear implementation/meaning, according to Step 3.
- For the fifth and sixth meanings, which are based on user profile and user-item dissimilarity, respectively, since they can be included by *difference*, we discard them (or merge them) according to Step 1.
- Finally, for the seventh meaning, item profile-based *novelty*, since it has no overlap with other disentangled factors (i.e., *random*, *unpopularity*, and *difference*), we take it as an independent factor *novelty* and assign it a clear meaning, according to Step 3.

The disentanglement results are shown in the right part of Fig. 2, where five factors are independent. The sum of meanings (together with that of interactive factors) after disentanglement remains the same as before.

Then we deal with seven factors in Fig. 3. The factors of *accuracy* and *relevance* have three meanings based on user-item similarity, user behavior, and user profile, respectively. In addition to these three bases, *usefulness* is also related to the user response. The factor of *value* consists of two parts, *usefulness* and the emotion-based value. The factors of *interestingness* and *insight* have two meanings based on user-item similarity and user profile, respectively; The factor of *attractiveness* is described based on the user profile; they are also included by *accuracy* and *relevance*.

Again, we treat non-interactive meanings with our disentanglement principles. The fourth and fifth meanings, which are based on user response or user emotion, are taken as interactive factors.

- For the first meaning, which is based on user-item similarity, since this meaning has no overlap with disentangled factors (which is null currently), we explore Step 3 of the three steps of disentanglement, i.e., take all the meaning and take it as an independent factor, *relevance*, and assign it a clear implementation/meaning as shown in the fourth column in Fig. 3.

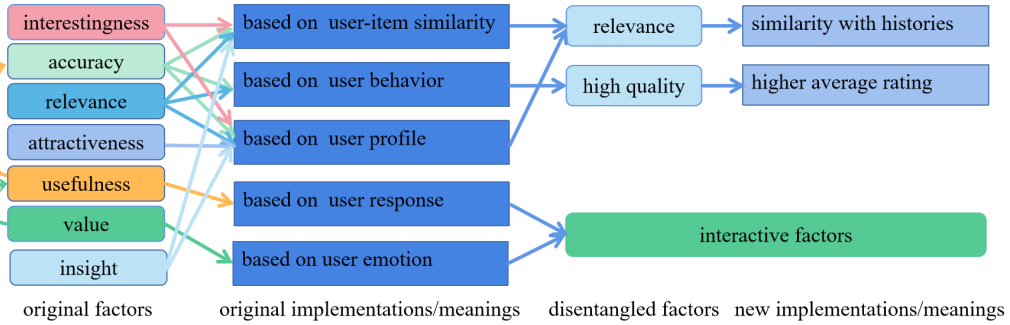


Fig. 3. Illustration of Factor Disentanglement: Factors for Matching User Interest.

- For the second meaning, which is user-behavior-based *accuracy*, since this meaning has no overlap with disentangled factors (i.e., *relevance*), we take it with a new factor of *high quality*, and assign it a clear implementation/meaning, according to Step 3.
- For the third meaning, which is based on user profile, since it can be included by the disentangled factors (i.e., *relevance*), we merge it according to Step 1.

The disentangled factors are shown on the right part of Fig. 3, where the factors are independent. Moreover, we take the others based on user response and emotion as interactive factors. Again, the sum of meanings after disentanglement remains the same as before. Note that we mainly study non-interactive factors with an objective data-driven approach. We will leave interactive factors to future studies.

3.3.2 Meaning Clarification. We clarify the meaning and possible implementations of each factor as follows:

For "relevance", "difference", and "diversity", we keep one meaning for each factor to maintain factor independence, that is, similarity for *relevance*, dissimilarity of the recommendation and historical items for *difference*, and dissimilarity between items in the recommendation list for *diversity*.

Factor 1: relevance. The factor of *relevance* is the main basis of *accuracy*, which is the goal of most recommendation methods. It could be measured in terms of the similarity between recommendations and user long-term preferences, short-term demands, and collaborative filtering. And the similarity metric depends on the application scenarios. For example, the distance between the nodes could be used to measure their similarity in the embedding space.

Factor 2: difference. Existing serendipity recommendation methods denote differences from user expectation, user profile, and items recommended by the primitive model as *unexpectedness* or *novelty*. We employ a more clear concept *difference* to represent the dissimilarity between recommendations and the historical items of the users.

Factor 3: diversity. There are two types of *diversity* in the literature: difference from the historical items of the user and difference among recommended items. We take the former as *difference* and the latter as *diversity*.

Factor 4: novelty. There are three implementations for "novelty" [39], new to the system (based on item profile), new to the user (based on user-item dissimilarity), and forgotten items (based on user profile). Here, we take the first as *novelty* and the other two as "difference", to maintain meaning coverage and factor independence.

Factor 5: unpopularity. Besides novelty, another factor could also help to recommend items unknown to users, in particular items in the long tail [81]. This factor is denoted as *unexpectedness* in [81]; for clarity, we employ a more direct concept *unpopularity* to denote it. It can be measured based on historical interactions (e.g., rating, purchasing, clicking).

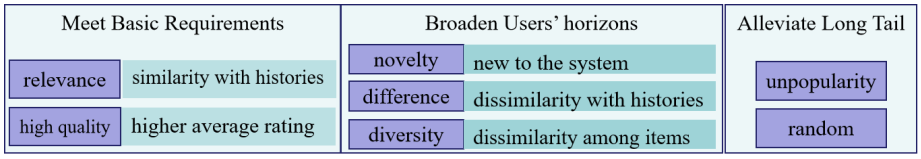


Fig. 4. A Brief Summary of Seven Factors.

Factor 6: high quality. This factor plays an important role in user decision-making. Items with higher ratings usually arouse user interest and improve satisfaction with a higher possibility. Some researchers even think that only items with higher ratings are relevant to users [30]. In this paper, items with a higher possibility to increase user satisfaction are regarded as high quality (e.g., higher ratings).

Factor 7: random. This factor is very special because for which all candidates have an equal chance of being selected. As for top-k recommendation for the target user u_t, k items are randomly selected to construct the recommendation list $R \subseteq C_{u_t}$.

Interactive Factors. Besides the above factors, there are some factors related to the user's decision, response, or emotion, which are denoted as interactive factors, as shown in Fig. 2 and Fig. 3. For instance, the *usefulness* of recommendations should be judged by users in [26], the *unexpectedness* is decided by participants in [11] and the item *value* might rely on emotional aspects in [9, 67]. Some works have deeply studied interactive factors with user surveys and gained useful findings (e.g., [13, 42, 87]). This paper mainly focuses on non-interactive factors involving less human intervention.

Based on the above factor investigation, we extract three personalized factors: *relevance*, *difference*, and *diversity*, and four non-personalized factors: *novelty*, *unpopularity*, *high quality*, and *random*. Among the seven factors, *relevance* and *high quality* can meet users' basic requirements and avoid unrelated or unsatisfying recommendations; *novelty*, *difference*, and *diversity* can help to widen users' horizons; while *unpopularity* and *random* can alleviate the cold start problem or long tail effect, and enhance the fairness of recommendations. Fig. 4 briefly summarizes the seven factors.

It is worth noting that the names and meanings of factors after disentanglement are uniform in different domains, while their impacts on serendipity may vary with different domains, for which we propose an experimental evaluation framework in the following section. Last but not least, the obtained set of seven factors can serve as a basis of recommendation serendipity, from which the recommendation service providers can select and combine some of them to construct serendipity in their domains; they can also add new factors for their special requirements according to the three steps of our factor disentanglement.

4 EXPERIMENTAL FRAMEWORK

We propose a general experimental framework for evaluating the impact of factors on serendipity in different domains. We first introduce the overall framework. Next, we provide an implementation to illustrate how to apply it in practice. We also propose a quantitative method to measure the degree of disentanglement between any two factors, providing a basis for factor selection and combination.

4.1 Overall Framework

The overall framework (Fig. 5) consists of two parts: optimization strategies and comparative performance evaluation. For each factor to explore, e.g., $factor_i$, we would employ a strategy to generate the corresponding recommendations R_i and evaluate its performance on serendipity with metric m_{ser} and those on all factors with metric $m_1, ..., m_n$. The

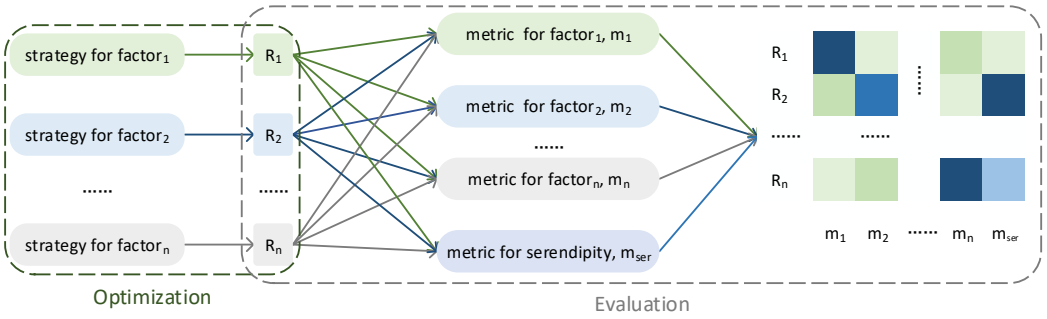


Fig. 5. Overview of Experimental Framework.

performance of recommendations on related factors would evaluate the effectiveness of the employed strategy, while the performance on serendipity can reflect the relative importance of each factor on recommendation serendipity. The framework can produce a list of factors ranked by their importance on serendipity in each domain.

4.2 An Implementation of The Experimental Framework

We propose a method to implement our framework. The recommendation strategy is used to optimize the corresponding factor, e.g., the strategy for "relevance" is expected to optimize the relevance of recommended items. As long as the strategy for each factor works as expected, we can then compare their effects on serendipity in the second part of our experimental framework. In this paper, we mainly check the relative importance of each factor in different domains, so we design the recommendation strategies independently.

4.2.1 Implementation of Recommendation. According to the results of factor clarification in Section 3.3, we implement a recommendation strategy and corresponding evaluation metric for each factor. We take the top-k recommendation as the basic framework and denote R as the set of recommended items, and C_{u_t} as the set of candidate items for target user u_t. For convenience, we also exploit the function $normalized(x)$ to convert the value into the range of [0,1].

Factor 1: Relevance. The factor of *relevance* denotes the similarity between recommendations and the labeled items I'_{u_t} in the test set. We employ a basic method of collaborative filtering to recommend relevant items based on similarity. While the similarity $sim(n_a, n_b)$ between two nodes n_a and n_b is calculated as the normalized dot product of their embedding vectors (www.cuemath.com/algebra/dot-product), in line with the embedding method in [33].

Strategy. The goal of *relevance*-based strategy is to maximize the similarity between the target user u_t and recommendations $R^{rel}_{u_t}$.

$$sim(n_a, n_b) = normalized(emb_{n_a} \cdot emb_{n_b}), \tag{1}$$

$$R^{rel}_{u_t} = \arg \max_{R \subseteq C_{u_t}} \sum_{i \in R} sim(u_t, i). \tag{2}$$

Metric. The average similarity of all recommended items generated by different recommendation strategies and the target user is taken as the evaluation metric, denoted as rel.

$$rel(R) = \frac{1}{|R|} \sum_{i \in R} sim(i, I'_{u_t}), \tag{3}$$

$$sim(i, I'_{u_t}) = \min_{i_a \in I'_{u_t}} sim(i, i_a). \tag{4}$$

Factor 2: Difference. The factor of *difference* represents the distance between recommendations and users' historical behaviors [52].

Strategy. For u_t, the recommendation set based on *difference* $R_{u_t}^{dif}$ is as follows:

$$dif(u_t, i_a) = 1 - \max_{i_b \in I_{u_t}} sim(i_a, i_b), \tag{5}$$

$$R_{u_t}^{dif} = \arg\max_{R \subseteq C_{u_t}} \sum_{i \in R} dif(u_t, i), \tag{6}$$

where I_{u_t} denotes items rated by u_t in the training set.

Metric. The factor of *difference* captures the dissimilarity among recommendations and historical items. So the metric dif is calculated as follows:

$$dif(R) = \frac{1}{|R|} \sum_{i \in R} 1 - sim(i, I_{u_t}). \tag{7}$$

Factor 3: Diversity. As mentioned in Section 3.3.2, the factor of *diversity* reflects the difference or dissimilarity among items.

Strategy. We employ the method in [14] to optimize the *diversity*. Suppose L is a kernel matrix that satisfies the condition, and L_R is the sub-matrix of L that is indexed by items in R. The possibility of recommending R is proportional to the determinant of L_R. The optimization objective is to find an item subset $R \subseteq C_{u_t}$ with the biggest determinant.

$$P(R) \propto det(L_R), \tag{8}$$

$$R_{u_t}^{div} = \arg\max_{R \subseteq C_{u_t}} det(L_R). \tag{9}$$

Metric. The factor of *diversity* denotes the difference among recommendations. To keep consistent with other metrics, the similarity between any two items in a list is exploited for measuring diversity here. It is worth noting that the above strategy is for optimizing list-wise diversity which is commonly pursued in diversity-oriented recommendation. While the metric exploits pair-wise similarity and measures the diversity of a whole list.

$$div(R) = \frac{1}{|R||R|} \sum_{i_a, i_b \in R} 1 - sim(i_a, i_b). \tag{10}$$

Factor 4: Novelty. For the factor of *novelty*, we regard items new to the system as novel ones, as mentioned in Section 3.3.2.

Strategy. We utilize the release timestamp $time(i)$ of item i, to measure its *novelty* $nov(i)$. The larger the timestamp, the larger the *novelty* is. For the target user u_t, top k items of candidates C_{u_t} with the biggest $nov(\cdot)$ would be offered. Formally, we generate *novelty*-based recommendation $R_{u_t}^{nov}$ as follows:

$$nov(i) = normalized(time(i)), \tag{11}$$

$$R_{u_t}^{nov} = \arg\max_{R \subseteq C_{u_t}} \sum_{i \in R} nov(i). \tag{12}$$

Metric. We employ the mean values of *novelty* scores of items in the recommendation set R to measure the *novelty* of recommendations.

$$nov(R) = \frac{1}{|R|} \sum_{i \in R} nov(i). \tag{13}$$

Factor 5: Unpopularity. The factor of *unpopularity* is another attractive factor to relieve the long-tail effect [81]. Items with fewer ratings tend to be more unpopular.

Strategy. Similar to [81], we employ the number of ratings to item i, $|rating(,i)|$, as the basic to measure its *unpopularity* $unpop(i)$. Formally, the *unpopularity*-based recommendation set $R_{u_t}^{unpop}$ for target user u_t, is calculated as follows:

$$unpop(i) = 1 - normalized(|rating(,i)|), \tag{14}$$

$$R_{u_t}^{unpop} = \arg\max_{R \subseteq C_{u_t}} \sum_{i \in R} unpop(i). \tag{15}$$

Metric. The *unpopularity* of recommendations $unpop(R)$ is the mean value of the *unpopularity* score of all items in R, as follows:

$$unpop(R) = \frac{1}{|R|} \sum_{i \in R} unpop(i). \tag{16}$$

Factor 6: High Quality. Items with *high quality* are more likely to gain user satisfaction, and thus receive high ratings [30].

Strategy. Similar to [30], we employ the average ratings of item i to measure its quality, denoted as $\overline{rating(,i)}$. Formally, the quality score $qua(i)$ of an item i and the *high quality*-based recommendation set $R_{u_t}^{qua}$, are calculated as follows:

$$qua(i) = normalized(\overline{rating(,i)}), \tag{17}$$

$$R_{u_t}^{qua} = \arg\max_{R \subseteq C_{u_t}} \sum_{i \in R} qua(i). \tag{18}$$

Metric. The quality of recommendations R is determined by the average quality of the recommended items.

$$qua(R) = \frac{1}{|R|} \sum_{i \in R} qua(i). \tag{19}$$

Factor 7: Random. The *random* factor is quite special because all candidates have an equal chance to be selected. Moreover, to compare the factor importance later, we must assign a score in [0,1) for each selected item.

Strategy. We first randomly assign a score $ran(i) \in [0, 1)$ to each candidate item $i \in C_{u_t}$ and then select the top-k items $R_{u_t}^{ran}$ according to the score, as follows:

$$ran(i) = random(), \tag{20}$$

$$R_{u_t}^{ran} = \arg\max_{R \subseteq C_{u_t}} \sum_{i \in R} ran(i). \tag{21}$$

Metric. Unlike other factor-optimizing strategies that try to optimize the corresponding factors, as long as the items are randomly selected, the *random*-based strategy can be taken as effective. Due to the special characteristic of the *random* factor, there is no metric to optimize the *random*-based strategy.

4.2.2 Implementation of Evaluation. We need to determine the serendipity metrics to check the factor impacts on recommendation serendipity. However, as mentioned before, there is no universally acknowledged definition or metric for serendipity up to now [72]. As pointed out by a recent survey in [108], there are three types of evaluation methods on serendipity: direct evaluation by a self-defined serendipity formula, indirect evaluation with its components, and evaluation based on user feedback.

To break the circle and keep the objective approach in mind (i.e., less human intervention), we search the literature and select two representative serendipity metrics in [26, 45, 52], denoted as $ser1$ and $ser2$, respectively.

Li et al. [51, 52] studied the serendipity 2018 dataset [42], and found that serendipity can be taken as the balance between *relevance* rel and *difference* dif; based on the two parts, they proposed $ser1$. While $ser2$ is based on a primitive model [26, 45], in which items generated by the primitive model R_P are taken as expected ones, and will be filtered out for generating serendipity recommendations. The level of serendipity is evaluated by the similarity between the remained items and the target item. Following the practice in [3, 101], we employ items with high average ratings or popularity as R_P. Formally, the calculations of $ser1$ and $ser2$ are as follows:

$$ser1(R) = \frac{2 * rel(R) * dif(R)}{rel(R) + dif(R)}, \tag{22}$$

$$ser2(R) = \frac{1}{|R - R_P|} \sum_{i \in R - R_P} sim(i, I_{u_t}). \tag{23}$$

The above two metrics evaluate serendipity in different ways. So they can work together to better check factor effects, that is, a factor that shows importance on both can be taken as really important. Furthermore, we also exploit two more recent serendipity metrics proposed by Fu et al. [23]. They labeled two datasets with serendipity by the crowdsourcing method and introduced two recall-based metrics, $HR\_ser@k$, and $NDCG\_ser@k$, defined based on the serendipity labels. $HR\_ser@k$ measures if any serendipity item is retrieved in the top-k position (1 for yes and 0 otherwise), while $NDCG\_ser@k$ measures both the existence and the positions of the serendipitous items in the top-k recommendation list. They are calculated as follows:

$$HR\_ser@k = \mathbb{I}(S \cap R \neq \emptyset), \tag{24}$$

$$NDCG\_ser@k = \sum_{i=1}^{k} \frac{serend\_score\_i(1 \text{ or } 0)}{log_2(i+1)}. \tag{25}$$

Where S represents the serendipity item set, and R represents the set of recommended items. $\mathbb{I}$ is the indicator function, which takes the value 1 if its argument is true, and 0 otherwise. This means that if at least one serendipity item appears in the first k positions of a recommendation list R, then the value of $HR\_ser@k$ for R is 1, and 0 otherwise.

Since the two metrics $HR\_ser@k$ and $NDCG\_ser@k$ need serendipity labels, they can only be tested in the two datasets labeled with serendipity by [23]. There are more metrics in the literature that can be flexibly selected and designed according to specific scenarios or requirements.

4.3 Degree of Disentanglement between Factors

We propose a new concept of the degree of disentanglement to measure how far two factors are being disentangled. This can help validate the effects of our factor disentangling method, as well as facilitate further understanding of factors and the combination of proper factors.

Intuitively, the more two factors are independent, the higher the degree of disentanglement between them, and the larger the distance between the recommended lists by the two factors' recommendation strategies. Therefore, we calculate it based on the distance of the recommendation results generated by different strategies. Let the recommendation lists produced by the two strategies of factors A and B be denoted as R_A and R_B, respectively. Without loss of generality, we exploit the normalized cosine distance to measure the distance between two items n_a and n_b in Eq. 26, where emb_{n_a} and emb_{n_b} are their embeddings. Based on this, we further define the distance from the recommendation list R_B to R_A,

i.e., $disent(R_A, R_B)$, as shown in Eq. 27. Note that this distance is asymmetrical, that is, the distance from R_B to R_A is calculated by selecting the items in R_B with the minimum distance to every item in R_A and outputting the minimum one, and vice versa. Some items may be selected several times if they are closer to items in the other list, while some other items may not be selected if they are further.

Different factors have different difficulties in disentangling. For instance, randomly generated recommendations tend to be more scattered, making it difficult to disentangle *random* and other factors. Thus, considering the varying difficulty of disentangling for different factors, we calculate the upper bounds for each factor's recommendation results and then normalize the degree of disentanglement based on their upper bounds. The details are as follows:

For a given recommendation list R, the calculation of the upper bound of its distant recommendation list is shown in Eq. 28. That is, we need to find an optimal subset R' from the candidate items C_{u_t} that maximizes the distance with R. However, this problem involves selecting the optimal subset and solving it for each user's recommendation would be computationally too expensive. To reduce the computation cost, we propose an alternative upper bound and rewrite the formula as Eq. 29. In this equation, R_F represents the recommendation lists generated by all strategies of factors defined in the experimental framework, with each list corresponding to each factor. This approach significantly reduces the search space for the upper bounds. Finally, we obtain the normalized and symmetric degree of disentanglement between two factors A and B by integrating the normalized distance from R_B to R_A and that from R_A to R_B, as in Eq. 30.

$$dis(n_a, n_b) = (1 - \frac{emb_{n_a} \cdot emb_{n_b}}{|emb_{n_a}| \cdot |emb_{n_b}|})/2, \tag{26}$$

$$disent(R_A, R_B) = \frac{1}{|R_A|} \sum_{i_a \in R_A} \min_{i_b \in R_B} \{dis(i_a, i_b)\}, \tag{27}$$

$$UpBound(R) = \max_{R' \in C_{u_t}} \{disent(R, R')\}, \tag{28}$$

$$UpBound(R) = \max_{R' \in R_F} \{disent(R, R')\}, \tag{29}$$

$$disent_{norm}(A, B) = \frac{1}{2}[\frac{disent(R_A, R_B)}{UpBound(R_A)} + \frac{disent(R_B, R_A)}{UpBound(R_B)}]. \tag{30}$$

Quantitatively measuring the degree of disentanglement between factors can help to quantify the information redundancy of any two factors, and guide the selection of the most proper factor combinations for serendipity in different domains, thus providing a novel perspective and theoretical support for understanding serendipity factors and optimizing serendipity recommendations.

5 EXPERIMENTAL EVALUATION

In this section, we apply our general framework and implementation in large-scale cross-domain datasets [5]. We have three goals for experiments: (1) to validate the effectiveness of the proposed recommendation strategies, by checking if they can optimize the corresponding factors, (2) to check the impacts of factors on recommendation serendipity by comparing their relative importance in each domain, and (3) to explore the ways of effective factor combination by measuring their degree of disentanglement and checking the effects of different combinations.

We verify the effectiveness of seven strategies for optimizing seven serendipity factors on six recommendation metrics (because the *random* factor is not taken as a recommendation metric) and four serendipity metrics (i.e., *ser1, ser2*, *HR_ser@k* and *NDCG_ser@k*), and explore what factors contribute more for serendipity, with the nine datasets in Table

[5]The code can be found at: https://github.com/csjwj2023/factors-of-serendipity-recommendation.

Table 1. Statistics of Datasets.

Dataset	#Users	#Items	#Interactions	aveInt/user	aveInt/item	Density
kindle	25927	58883	662994	25.57	11.26	0.043%
electronics	49072	82932	835578	17.03	10.08	0.021%
home	1395	1171	25445	18.24	21.73	0.016%
clothing	13058	62137	185551	14.21	2.99	0.023%
tool	1173	629	22633	19.29	35.98	3.068%
beauty	1197	693	26983	22.54	38.94	3.253%
sport	10849	35368	172241	15.88	4.87	0.045%
ser_bk	2346	113876	265037	112.97	2.33	0.099%
ser_mv	621	23952	74970	120.72	3.13	0.504%

1. For $ser2$, we employ 100 most popular items and 100 items with the highest ratings as primitive recommendations, since 100 is sufficient for all users in our datasets, and it is also believed to be enough in many domains. Therefore, the recommended items filtered out by the primitive model are quite different from those on $ser1$. In this way, we can better understand serendipity from different perspectives. Note that the *high quality* factor is not considered for $ser2$ because high-quality items that have the highest ratings are excluded in $ser2$.

Rationality of Serendipity Metrics. The four serendipity metrics evaluate the serendipitous degree from different perspectives. The two metrics $ser1$ and $ser2$ involve little human intervention. While $HR\_ser@k$ and $NDCG\_ser@k$ are recall-based metrics that involve human-labeled data. $ser1$ belongs to the category of component-based serendipity metrics, which exploits the results of the two components of *relevance* and *difference*. It reflects the common fact that serendipity comes from difference, meanwhile, with some basic relevance. While $ser2$ belongs to the category of formula-based serendipity metrics, which evaluates the similarity between the recommended items that have been filtered by the primitive model and the target item. Therefore, we can say that the four metrics generally cover the most understanding of recommendation serendipity.

5.1 Experimental Settings

5.1.1 Datasets and Basic Settings. We conduct experiments on nine representative datasets, which mainly come from Amazon review data (Amazon Dataset). They cross various domains and have different statistical characteristics. Moreover, we also take into account the impact of user responses on serendipity and conduct experiments on two serendipity datasets provided by Fu et al. [23]. These two datasets provide large-scale labeled data on serendipity, obtained through crowdsourcing methods that process user comments in Amazon books and movies.

We preprocess data following the approach of CLSR [102] and SUM [53], which adopt an iterative filtering approach to implement the 10-core setting to ensure data quality. It executes two iterations: first, delete items with fewer than ten interactions; second, retain users with at least ten interactions and remove the others. Note that after the second iteration, some items may retain fewer than ten interactions because some of their reviewers were deleted. The statistics are shown in Table 1, and the detailed distribution of interactions per item is shown in Table 2.

For the training and testing set division, [37] pointed out that it should follow the global timeline to avoid data leakage. We follow a commonly used setting: according to the timestamps of each user's interaction sequence, we divide each dataset with the first 80% interactions of each user as the training set and the other 20% as the testing set. We conducted all the experiments ten times and checked their average performance.

For the accuracy and the efficiency of the similarity calculation, we generate representations of users and items with a recent representation method particularly designed for recommendation system, LightGCN [33], which has been

Table 2. The Distribution of Interactions Per Item in Nine Datasets.

Dataset	1	2	3	4	5	6	7	8	9	10	>10
kindle	9.67	8.38	7.77	7.07	6.81	6.33	6.08	5.68	5.18	4.33	32.7
electronics	19.76	16.53	12.16	8.9	6.5	4.82	3.68	2.97	2.42	1.98	20.28
home	0	0	0	0	0	0	0	0	0	9.14	90.86
clothing	42.68	23.45	11.97	6.79	3.95	2.57	1.8	1.28	0.94	0.82	3.75
tool	0	0	0	0	0	0	0	0	0	0	100
beauty	0	0	0	0	0	0	0	0	0	0	100
sport	34.61	19.5	12.05	7.64	5.15	3.75	2.59	2.14	1.78	1.41	9.38
ser_bk	58.27	19.01	7.95	4.4	2.69	1.81	1.33	0.92	0.69	0.5	2.43
ser_mv	48.4	19.97	9.11	5.6	3.42	2.67	1.72	1.63	1.1	1.02	5.36

verified to be effective in recommendation tasks. We conduct top-k recommendations and check the performance with k= 5, 10, 15, and 20. The results show similar trends. Here we mainly show the results with $k = 20$ to avoid redundancy.

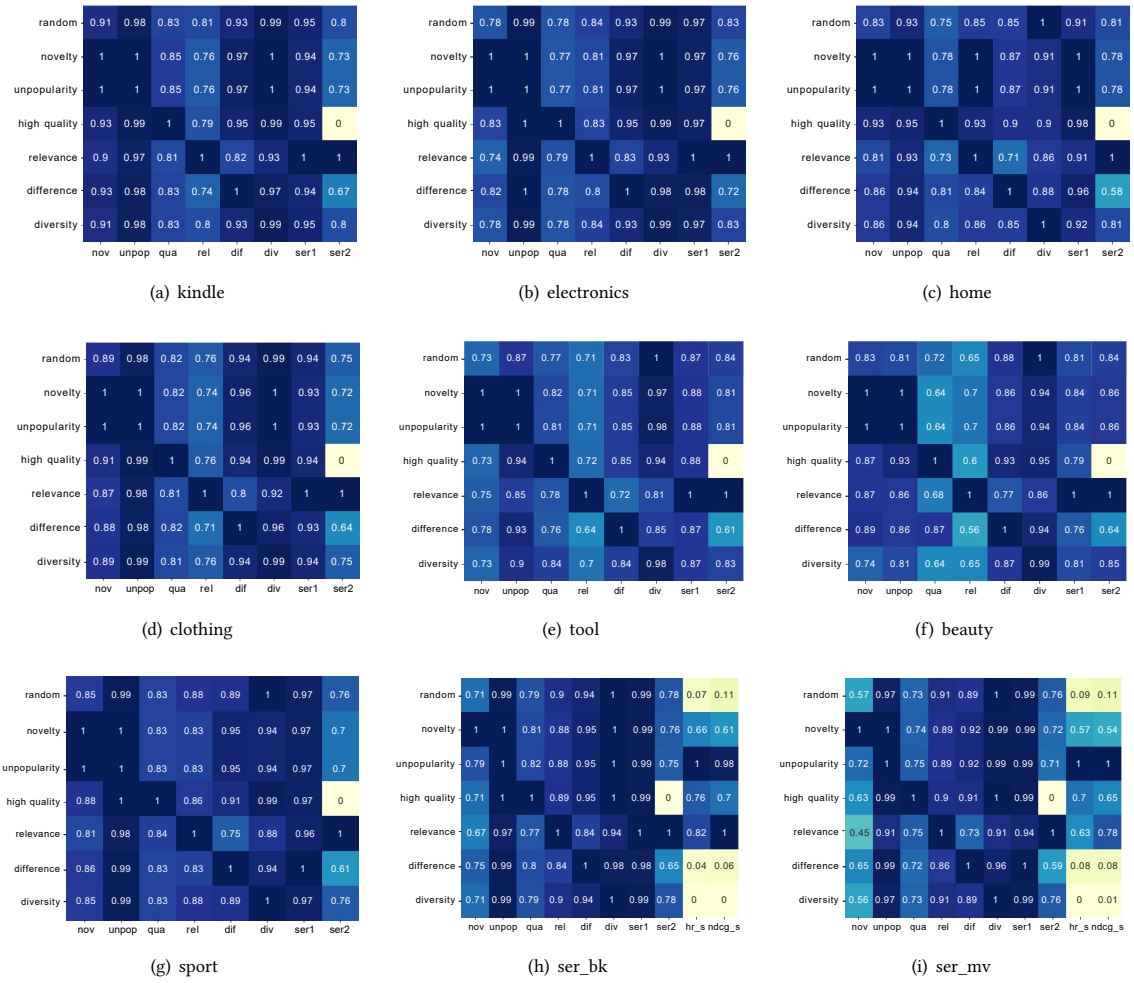
5.1.2 Candidate Generation. For the efficiency of the evaluation process and the reliability of results, we generate candidates with ten random seeds and conduct experiments on them. There are two reasons for this setting. First, it costs lots of time to run the seven strategies on the nine original whole datasets. Second, the goal of performance evaluation is not to obtain the exact values but to compare the relative importance of factors. Hence, we choose to generate candidates and we try to keep similar statistical characteristics to that of the global. For each user u, we employ the proportionate stratified sampling strategy [83] to sample α unrated items(we set α as 1000 which is large enough). After that, we checked the statistics on the similarity of unrated users and items on the global and the candidate sets, and they show very similar statistical characteristics. This validates that the candidate sets are reasonable, and it is sufficient to conduct experiments on the candidates.

5.2 Effectiveness of Strategy

Fig. 6 displays the evaluation results on nine datasets. We take Fig. 6(a) as an example to describe the meaning, which displays the results on the kindle dataset. In this figure, we directly display the average score of all users on eight metrics (i.e., *nov*, *unpop*, ..., $ser1$, and $ser2$) of each factor's recommendation strategy. For a clear comparison, we fill the color of each unit with scores in each column. A deeper color indicates a greater contribution of the factor to the metric.

We can see that the corresponding strategy usually performs the best on the corresponding metrics, i.e., the factor of $novelty$ gets the highest score on the *nov* metric, as well as $unpopularity$ on *unpop*, $high\ quality$ on *qua*, $relevance$ on rel, and *difference* on dif. Therefore, we can say that most recommendations show a clear tendency for the corresponding factor and a relatively poor optimization for the rest. This indicates that **most strategies work as expected to optimize the corresponding factor.** Therefore, the strategies are validated to be effective.

There is only one exception, *diversity* strategy performs the best on the *div* metric in four datasets (home, sport, ser_bk, and ser_mv), and it ranks the second best in the other five datasets. We analyze the reason and find that in our framework (and in most existing diversity-oriented recommendation strategies), the employed method optimizes for set-wise *diversity*, that is, all items in the recommendation list are expected to be various. Whereas the metric *div* is for pair-wise *diversity*, that is, any two items in the recommendation list are expected to be less similar to each other. It is worth noting that the choice of pair-wise *div* metric is also reasonable because it can reflect the diversity from the perspective of individual users and items. While evaluating set-wise diversity usually needs some additional information like the category or aspect coverage. Additionally, the method utilizes a greedy algorithm to address an NP-hard problem, which may involve some approximation error.



(a) kindle (b) electronics (c) home

(d) clothing (e) tool (f) beauty

(g) sport (h) ser_bk (i) ser_mv

Fig. 6. Comparison of Various Strategies on Nine Datasets.

5.3 Impacts of Factors on Serendipity

In this section, we deeply analyze the impacts of factors on serendipity from multiple perspectives: performance on four serendipity metrics, performance in different datasets/domains, and possible impacts of datasets/domains.

5.3.1 Performance on Four Serendipity Metrics. In this subsection, we check the overall factor importance on the four serendipity metrics: *ser*1, *ser*2, *HR_ser@k*, and *NDCG_ser@k* (Section 4.2.2), as shown in Fig. 6. Due to space limitations in the last two subfigures, we use *hr_s* and *ndcg_s* for short of *HR_ser@k* and *NDCG_ser@k*, respectively. The main findings are as follows:

Four factors of *unpopularity*, *relevance*, *high quality* and *novelty* make more impacts on all four serendipity metrics. Since serendipity bears a subjective nature, besides 7 review datasets without human intervention on serendipity, we also conduct experiments on two large-scale serendipity datasets labeled by humans with a crowdsourcing approach [23], ser_bk and ser_mv. The last two columns of Figs. 6(h) and 6(i) show that on *HR_ser@k* and

NDCG_ser@k, the most important factors include *unpopularity*, *novelty*, *high quality* and *relevance* in ser_bk and ser_mv. The reason may be that humans usually prefer new items, which increases the chances for cold and new items to be evaluated as positive serendipity samples. Meanwhile, the four factors also show relatively high importance on $ser1$ and $ser2$ (except for high-quality, which is excluded on $ser2$) in the two labeled datasets. Therefore, we can say that the performance on the four serendipity metrics together reflects the importance of those four factors.

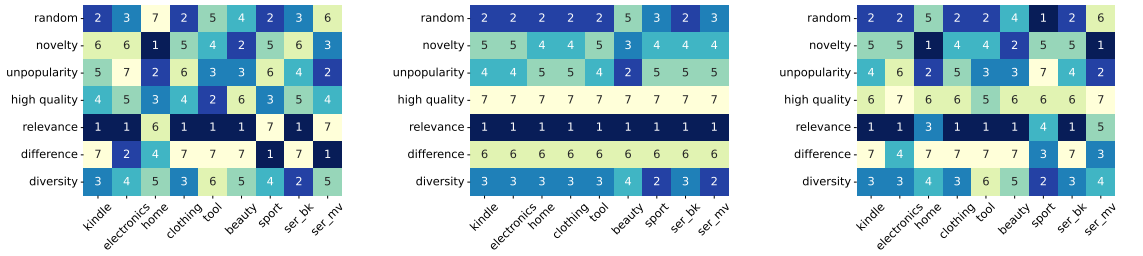
Other three factors of *random*, *difference*, and *diversity* make less impacts on *HR_ser@k* and *NDCG_ser@k*, because of the sparsity of serendipity labels. In Figs. 6(h) and 6(i) it can also be observed that the other three factors, *random*, *difference*, and *diversity*, gain much less importance on *HR_ser@k* and *NDCG_ser@k*. While they gain higher importance on $ser1$ and $ser2$. This may be because the serendipity labels are quite sparse in the two datasets (1% [23]), and each user usually has a few positive samples and multiple negative samples on serendipity. Therefore, it is difficult for the recommendation strategies based on *random*, *difference*, and *diversity* to hit the positive labels, leading to poor performance on *HR_ser@k* and *NDCG_ser@k*. It is worth noting that manual annotation requires additional costs and may introduce subjective biases. Therefore, we suggest integrating metrics that do not require additional manual labeling, like $ser1$ and $ser2$.

Furthermore, in Fig. 6 the factors' importance on $ser1$ and $ser2$ are quite different in nine datasets, just as expected when designing the two metrics (Eqs. 22 and 23). We will integrate the performances on the two metrics to comprehensively analyze factor importance in different domains.

5.3.2 Performance in Different Datasets/Domains. We summarize the serendipity rankings of $ser1$ and $ser2$ on all datasets in Fig. 7 to compare the factor impacts directly for serendipity on various domains. Fig. 7(a) and Fig. 7(b) display the rankings of factors on nine datasets in terms of $ser1$ and $ser2$, and Fig. 7(c) shows the overall tendency by summing the factor rankings of $ser1$ and $ser2$, denoted as ser-combined. The more a factor contributes to serendipity, the higher it ranks. Meanwhile, the higher the ranking, the darker the unit color is. Taking the first column in Fig. 7(a) as an example, it shows that on kindle, *relevance* ranks first, indicating that it contributes the most to $ser1$. The ranks of other factors are as follows: *random, diversity, high quality, unpopularity, novelty*, and *difference*. Moreover, the orders are different in different columns, indicating that the roles of factors on serendipity vary with datasets or domains.

In general, *relevance*, *diversity*, and *random* make more impacts on serendipity. We count the number of datasets on which each factor ranks within the top 4 of all seven factors, according to the combined serendipity in Fig. 7(c). The first four factors are *relevance* (8 times), *diversity* (7 times), *random* (7 times), and *unpopularity* (6 times). Meanwhile, *novelty* (5 times) rank in the middle, *difference* and *high quality* rank at the bottom. Since *random* strategy has some uncertainty, we suggest to exploit it more carefully. Therefore, with the cross-domain experiments, we believe *relevance*, *diversity* and *unpopularity* are beneficial serendipity factors for future research. Moreover, the higher importance of *unpopularity* also indicates that serendipity recommendation is beneficial for cold start items with fewer ratings and can help alleviate the long tail effects.

Factor rankings on $ser1$ vary among different datasets, while those on $ser2$ are quite consistent with most datasets. As shown in Fig. 7(a), the rankings of seven factors on $ser1$ show much variance in nine datasets. For instance, to observe along the rows, *relevance* ranks first on six datasets, while it ranks 6th or 7th on home, sport, and ser_mv. Meanwhile, *difference* ranks last on five datasets, while it ranks higher on electronics, home, sport, and ser_mv. Moreover, the other five factors usually rank in the middle, with rankings varying from 2nd to 6th. Next, to observe along the columns in Fig. 7(a), factor rankings on some datasets share a similar trend, like kindle and ser_bk. On the two datasets, *novelty*, *relevance*, and *difference* share the same rankings (i.e., *novelty* ranks 6th, *relevance* ranks 1st, and



(a) $ser1$ (b) $ser2$ (c) ser-combined

Fig. 7. Rankings of Factor Importance on (a) $ser1$, (b) $ser2$ and (c) ser-combined.

difference ranks last, respectively), and other factors' rankings differ no more than one. While the seven factors have the same rankings on *ser_combined* in the two datasets, as shown in Fig. 7(c).

As for $ser2$, Fig. 7(b) shows that *relevance* consistently ranks first and *difference* ranks 6th on $ser2$, while *high quality* consistently ranks last because it is excluded as primitive results, and it is non-personalized which leads to low relevance with the target user. Hence, we suggest exploiting *high quality* and *relevance* (or other personalized factors) together in practice. Moreover, *random* and *diversity* rank 2nd or 3rd, *novelty* and *unpopularity* rank 4th or 5th in most datasets. The reason may be that $ser2$ is mainly based on the similarity between items (Eq. 23), which is close to the *relevance*.

We analyze the difference between performance on $ser1$ and $ser2$ and find two possible reasons: (1) The design of the two metrics is quite different from each other. $ser1$ is based on the balance of *difference* and *relevance* between recommended and real interacted items. While $ser2$ is based on the recommendation, excluding that from a primitive model. (2) The importance of *relevance* is different from that of *difference* among nine datasets, which leads to the variance on $ser1$.

Considering $ser1$ and $ser2$ together, *unpopularity* and *novelty* have similar rankings in all the nine datasets. In Fig. 7(a) and Fig. 7(b), although the rankings of *unpopularity* and *novelty* vary with datasets, they have similar rankings in the same dataset; In most datasets, their difference is no more than one; only in ser_bk, their ranking difference is 2. We will analyze the possible reasons from the perspective of factor correlation in Section 5.4.

5.3.3 Possible Impact of Datasets/Domains. To find the possible impact of datasets or domains on the importance of serendipity factors, we further analyze item and user distribution in the embedding space. We calculate the mean value μ and standard deviation value σ of the distances between user-item and item-item in the training and testing sets, as shown in Tables 3 and 4, respectively. Without loss of generality, we use the normalized cosine distance $dis(n_a, n_b)$ to measure the distance of two items as in Eq. 26. We also list the average interactions per user and item in Table 1, and the distribution of interactions per item in Table 2. Note that our data preprocessing only guarantees that each user has at least ten interactions, while only in home, tool, and beauty, most items have at least 10 interactions. In the other six datasets, most items have less than 10 interactions.

Tables 3 and 4 consist of user-item distance and item-item distance. Each of the two distances consists of three parts, indicating the performance (i.e., the mean value μ and standard deviation value σ) on the rated items (i.e., the first part), on all items (i.e., the second part), and the ratio of rated over all items (i.e., the third part). The items rated by a user can be taken as related to him. The ratio depicts the relative aggregation level of related user-item representations. In general, a smaller ratio indicates that related nodes are more aggregated and users focus more on historical interest, while a larger ratio indicates a broader interest distribution. We gain several findings, as follows:

Table 3. Statistics on Distance in Embedding Space (in the Training Set).

dataset	user-item distance						item-item distance					
	rated		all		rated/all		rated		all		rated/all	
	μ_{ui}	σ_{ui}	μ_{ui}	σ_{ui}	μ_{ui}	σ_{ui}	μ_{ii}	σ_{ii}	μ_{ii}	σ_{ii}	μ_{ii}	σ_{ii}
kindle	0.243	0.107	0.504	0.066	0.481	1.631	0.382	0.113	0.219	0.250	1.745	0.452
electronics	0.286	0.127	0.510	0.066	0.562	1.921	0.416	0.125	0.295	0.250	1.408	0.500
home	0.292	0.107	0.499	0.096	0.585	1.114	0.404	0.146	0.500	0.099	0.809	1.477
clothing	0.193	0.166	0.510	0.071	0.379	2.335	0.310	0.173	0.106	0.208	2.916	0.830
tool	0.318	0.098	0.493	0.103	0.646	0.957	0.405	0.144	0.498	0.098	0.813	1.464
beauty	0.335	0.099	0.503	0.084	0.666	1.178	0.422	0.134	0.499	0.087	0.846	1.553
sport	0.199	0.112	0.510	0.085	0.391	1.325	0.327	0.140	0.493	0.096	0.663	1.461
ser_bk	0.189	0.111	0.501	0.081	0.376	1.368	0.357	0.124	0.010	0.073	34.098	1.716
ser_mv	0.236	0.125	0.502	0.075	0.471	1.668	0.375	0.119	0.014	0.084	25.922	1.405

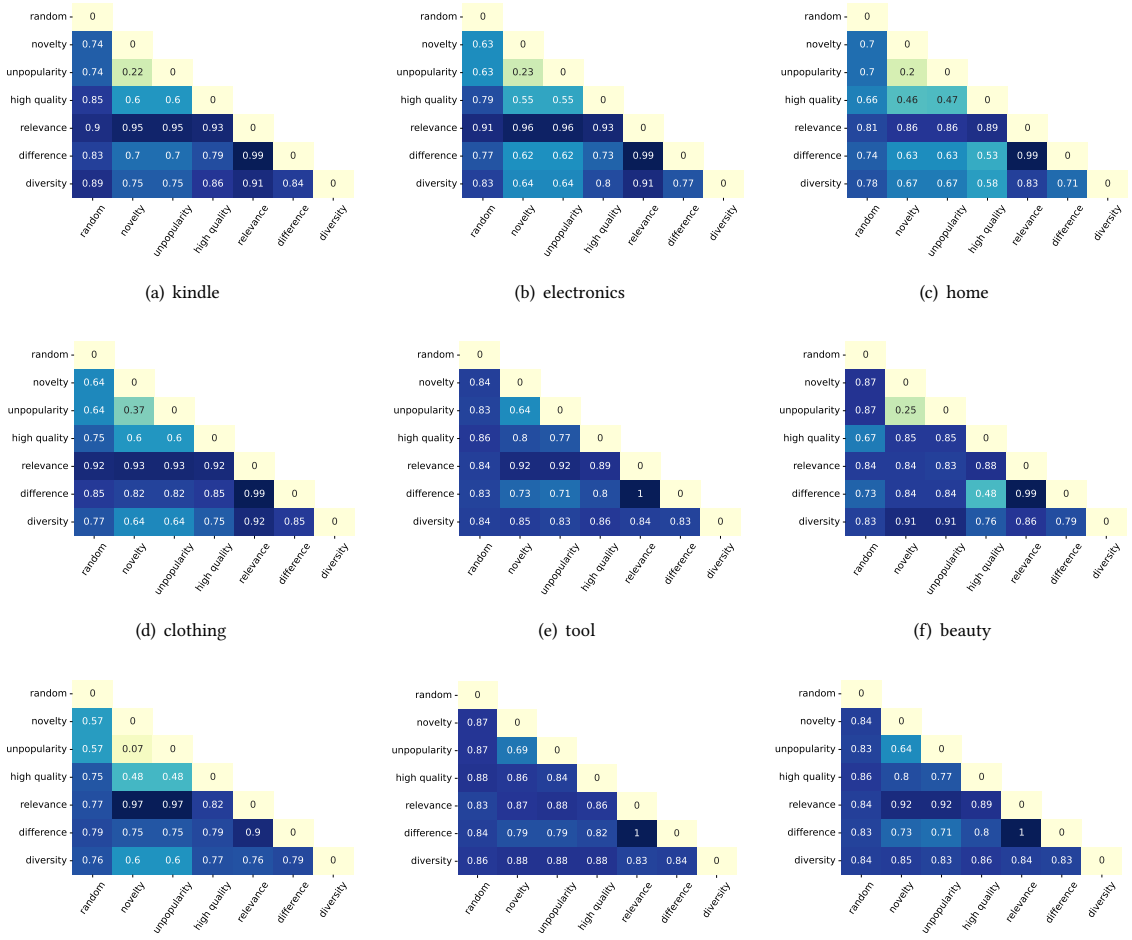
Table 4. Statistics on Distance in Embedding Space (in the Testing Set).

dataset	user-item distance						item-item distance					
	rated		all		rated/all		rated		all		rated/all	
	μ_{ui}	σ_{ui}	μ_{ui}	σ_{ui}	μ_{ui}	σ_{ui}	μ_{ii}	σ_{ii}	μ_{ii}	σ_{ii}	μ_{ii}	σ_{ii}
kindle	0.347	0.139	0.504	0.066	0.688	2.121	0.371	0.155	0.219	0.250	1.696	0.620
electronics	0.373	0.149	0.510	0.066	0.732	2.248	0.367	0.197	0.295	0.250	1.244	0.787
home	0.441	0.165	0.499	0.096	0.882	1.725	0.326	0.223	0.500	0.099	0.652	2.254
clothing	0.273	0.204	0.510	0.071	0.535	2.856	0.292	0.228	0.106	0.208	2.744	1.095
tool	0.323	0.103	0.493	0.103	0.654	0.999	0.324	0.192	0.498	0.098	0.651	1.952
beauty	0.318	0.102	0.503	0.084	0.632	1.212	0.342	0.182	0.499	0.087	0.686	2.098
sport	0.444	0.103	0.510	0.085	0.870	1.214	0.356	0.217	0.493	0.096	0.723	2.272
ser_bk	0.415	0.110	0.501	0.081	0.827	1.359	3.83e-8	3.86e-8	0.010	0.073	2.31e-7	5.33e-7
ser_mv	0.450	0.108	0.502	0.075	0.896	1.435	7.27e-9	3.83e-8	0.014	0.084	5.03e-7	4.53e-7

In general, user interests' distance remains stable in the two datasets of tools and beauty, while it becomes larger with time in the other seven datasets. Table 3 displays the statistics of the first 80% interactions (i.e., the training set), and Table 4 displays the last 20% interactions (i.e., the testing set). Comparing the values of μ_{ui} (the 2nd column, user-item distance, rated) in Table 3 with those in Table 4, we can observe the rather significant changes in seven datasets other than tool and beauty. For instance, $\Delta\mu_{ui}$ for tool and beauty is $0.323 - 0.318 = 0.005$ and $0.318 - 0.335 = -0.017$ respectively; while $\Delta\mu_{ui}$ for kindle and ser_mv is $0.347 - 0.243 = 0.104$ and $0.450 - 0.236 = 0.214$, respectively. It suggests that users' interests may be more stable in the two domains of tools and beauty, while they evolve much in the other seven domains. We further analyze the reason and find that, there are many more items and fewer average interactions per item in the other seven domains (as shown in the third column "Items" and the sixth column "aveInt/item", in Table 1), indicating more choices and more variance.

The role of factors in affecting serendipity varies with datasets, and the domain features also matter. Fig. 7 shows that the rankings on different factors vary with datasets. Combining Fig. 7(c) with Tables 3 and 4, we observe that datasets with similar statistical characteristics and similar domain features also exhibit similar factor ranks on serendipity. For instance, the rankings of factors in the kindle and ser_bk datasets are the same on *ser*-combined, while the mean values of the two datasets are relatively small, and the two involve similar domains: kindle is for electronic reading and ser_bk is for reading in Amazon books.

A similar finding is also observed in home and ser_mv, as well as tool and beauty, where the mean values of user-item distances are close to each other in both the training and testing datasets. Taking home and ser_mv for instance, among



(a) kindle (b) electronics (c) home

(d) clothing (e) tool (f) beauty

(g) sport (h) ser_bk (i) ser_mv

Fig. 8. The Degree of Disentanglement of Various Strategies on the Nine Datasets.

all the seven factors, the rankings of *novelty*, *unpopularity*, and *diversity* are the same, and two other factors (i.e., *random* and *high quality*) differ in ranking by only one, indicating some consistency between the factor rankings and the statistical characteristics.

In summary, the statistical characteristics of datasets can reflect the overall features of user interactions in different domains, which will further impact factor importance on recommendation serendipity. This finding is very helpful for future research in that we can guide the strategy design with the domain features. For instance, if users' interest is more focused in some domain, then *relevance* can be given more weight, and vice versa.

5.4 Correlation among Serendipity Factors

Fig. 6 also reflects the correlation among factors. Similar to the rankings in Fig. 7, we can also check the factor importance on the other six metrics, i.e., *nov*, *unpop*, *qua*, *rel*, dif, and *div*. Taking the *nov* metric for instance, besides *novelty* itself, *unpopularity* and *difference* contribute more than the others. To further study the correlation among serendipity factors, we calculate the degree of disentanglement among any two factors using Eq. 30. Fig. 8 illustrates the degree of

disentanglement between recommendations generated by the strategy corresponding to different factors in the nine datasets. We have the following findings:

All the degrees of disentanglement are quite large between any two factors. This validates the effectiveness of our factor-disentanglement approach and indicates that different factors can measure serendipity from different perspectives.

The degree of disentanglement between *difference* and *relevance* is consistently high in all the nine datasets. We analyze the reason and find that, according to the definitions in Section 3.3.2, *difference* tends to recommend items that are different from the user's historical behavior, while *relevance* is more focused on the user's historical interests, representing the other side. Therefore, their recommendations exhibit significant differences. Based on this finding, we can combine the two factors for better serendipity recommendations.

The degree of disentanglement between *unpopularity* and *novelty* is consistently low in all the nine datasets. For *unpopularity* and *novelty*, the former is calculated by the normalized rating numbers, and the latter is calculated by the normalized time when an item is put into the system (i.e., new to the system). Since new items typically have fewer ratings than those that have existed longer, it is natural that the two strategies based on these factors often recommend similar items. However, they are not always consistent. For instance, some longer-existing items may also have higher unpopularity if they received very few ratings, while those items have a lower novelty.

Moreover, as shown in Fig. 6, both *novelty* and *unpopularity* achieve the best performance in each other's metrics in the seven general datasets, and second best (with itself performing the best) in the two labeled datasets. In addition, their rankings in serendipity are similar on nine different datasets, as shown in Fig. 7. This indicates that their correlation is higher than others, which is consistent with the relatively lower degree of disentanglement. This further validates the effectiveness of our factor disentangling approach.

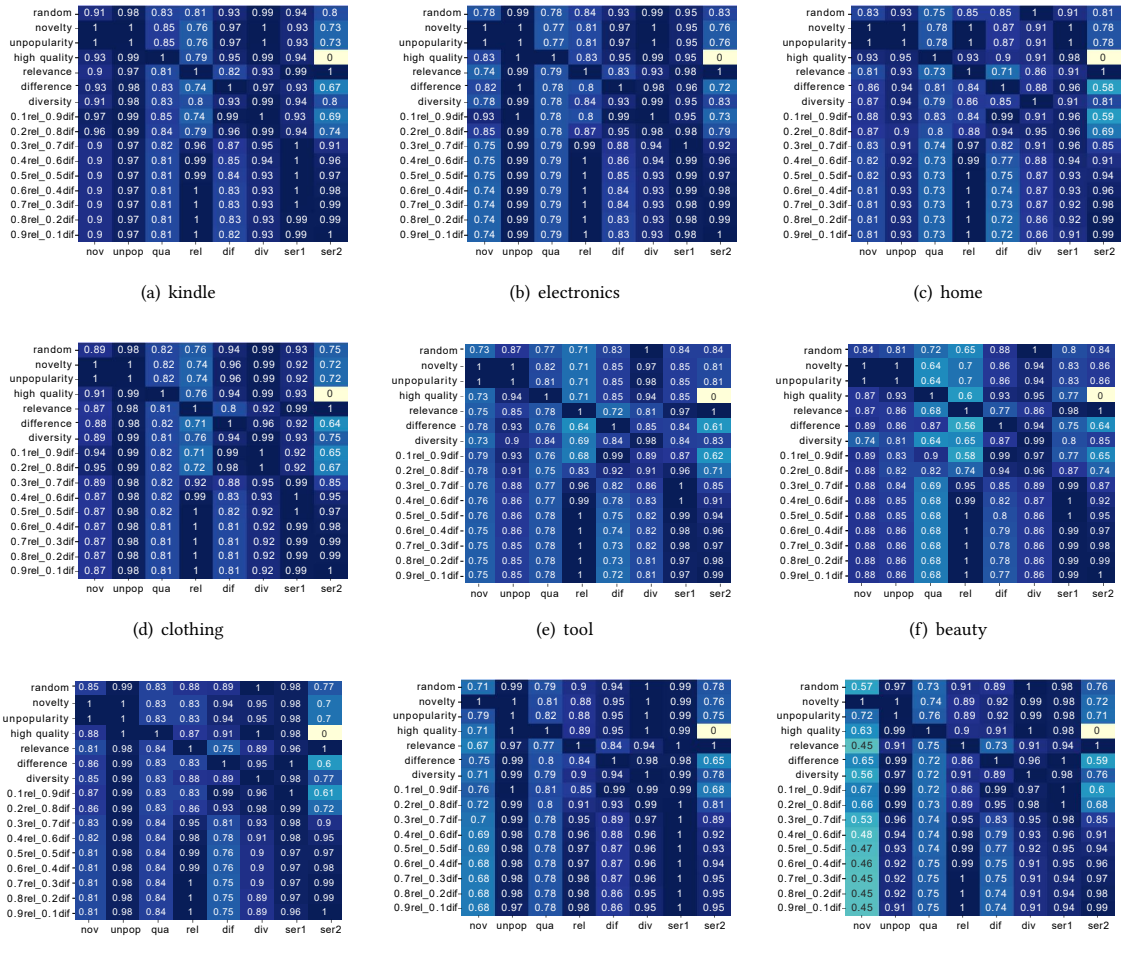
The degrees of disentanglement between *relevance* and all other factors are generally high in most datasets. This indicates two things: first, the recommendations by *relevance* based strategy are quite different from other factor-based strategies; second, it will bring benefits on serendipity when combining *relevance* and other factors since they usually produce different recommendations.

In summary, the degree of disentanglement can help understand the level of information redundancy of different factors and guide the selection of the most important factors for serendipity in different domains.

5.5 Combination of Serendipity Factors

In this section, we explore the impact of different factor combinations on recommendation serendipity. As analyzed in Section 5.4, the degree of disentanglement between *difference* and *relevance* is consistently high in all datasets, which indicates they exhibit the minimum redundancy of information. Therefore, we take the two factors as an example to study the effects of factor combination. We simply summarize the scores of the two factor-based strategies with linear weights from 0.1 to 0.9, and check the importance of the nine combinations, as illustrated in Fig. 9. For instance, $0.1rel\_0.9dif$ represents the combination of taking the score of relevance-based strategy with a weight of 0.1 and that of difference-based strategy with 0.9. We find that with the growth of relevance in the combination, the importance of the combination factor on $ser2$ continuously increases. As for $ser1$, there are two kinds of trends in the nine datasets, as follows:

When *difference* weighs more than *relevance* on $ser1$, the performance of all the nine combinations falls in between that of the two single factors. As shown in the seventh column in Figs. 9(c), 9(g), and 9(i), the importance of *difference* (i.e., the sixth row in the figures) is larger than that of *relevance* (i.e., the fifth row) on $ser1$ in the three domains



(a) kindle (b) electronics (c) home

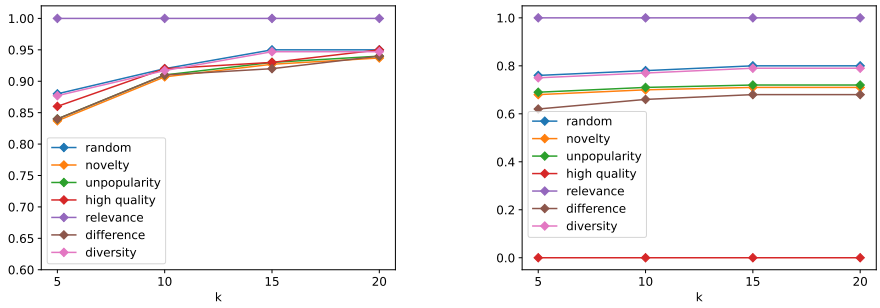
(d) clothing (e) tool (f) beauty

(g) sport (h) ser_bk (i) ser_mv

Fig. 9. Comparison between Combined Factors and Single Factor based Strategies on Nine Datasets.

of home, sport, and ser_mv. Then, the performance of the nine combinations from $0.1rel\_0.9dif$ to $0.9rel\_0.1dif$ decreases gradually on $ser1$. Together with the factor rankings of home, sport, and ser_mv in Fig. 7(a), it indicates that in the domain of home, users prefer new and unpopular products (i.e., *novelty* and *unpopularity* rank the highest). Meanwhile, in the domain of sport, *difference* and *high quality* are more important. In the domain of movie watching (i.e., ser_mv), users prefer more movies that are different from historical records, while *unpopularity* and *novelty* also rank higher.

When *relevance* weighs more than *difference* on $ser1$, the performance of all combinations has two stages, it first falls in between that of the two single factors, then it outperforms any single strategy. This can be seen in the seventh column in Figs. 9(a), 9(b), 9(d), 9(e), 9(f), 9(h). We attribute this to the combination of *difference* and *relevance*, which allows for the selection of relevant items at an appropriate distance from the user's historical interests. This approach effectively balances relevance while satisfying the user's exploratory needs. As shown in the sixth column in Table 4, the statistical characteristics of these datasets, i.e., the rated/all ratios of user-item distances μ_{ui}, are generally smaller than the other three datasets (i.e., home, sport, and ser_mv). It demonstrates that users' interests



(a) ser1 (b) ser2

Fig. 10. The Impacts of Different k, Taking the kindle Dataset as An Example.

are more concentrated. The factor rankings in Fig. 7(a) also validate that in these domains, users prefer items similar to historical records, where *relevance* ranks first. Therefore, the fusion approach that combines relevance can achieve performance on $ser1$ beyond any single factor. This allows for trade-offs between different factors of serendipity to accommodate the specific needs of different domains.

5.6 Impact of Possible Implementations and Parameters

Possible Implementations of Experimental Framework. In the general framework, it is flexible to select suitable ways to optimize factors, evaluate the performance of serendipity and its factors, as well as deal with new factors or combined factors. For instance, the recommendation can exploit personalized or non-personalized strategies, in either static or interactive environments; the performance evaluation can be measured with quantifiable metrics or user surveys. Furthermore, when applied to specific platforms, it only needs to implement factor-optimization strategies and serendipity metrics based on its own data processing and recommendation framework.

Note that although the absolute value of factor impacts in a domain may be different with different implementation strategies, the relative importance of factor rankings is stable in each domain, as long as the factor-based strategies can effectively optimize the corresponding factors and the experiments are conducted with consistent evaluation metrics and datasets. This is because, for a given target user, a given optimization metric, and a given dataset, the best k candidates that meet each optimization metric should be similar or even the same.

Possible Effects of Embedding Method. The values in Tables 3 and 4 may vary with different representation models, because they may generate different embeddings for users and items. Since LightGCN has been verified to be effective in recommendation tasks, we exploit it as the representation model. We also try some other embedding methods like word2vec [63]. While the absolute values are different, the relative trends are similar to those in Tables 3 and 4. Moreover, when applied to real platforms, it can directly use the original data processing (including the embedding method) in the platforms.

Impacts of Parameter k in top-k Recommendation. To ensure the stability of our results, we conduct the experiments with different k. We check the performance with k = 5, 10, 15, and 20 on the nine datasets, and the results show similar trends (see that for kindle in Fig. 10). Under the $ser1$ and $ser2$ metrics, as k increases, the performance of the strategies does not change significantly except for *relevance* and *diversity*. The reason is that the coverage rate of the recommendation increases as k increases, which leads to an increase in diversity. Other factors remain with the same relative relationship.

6 CONCLUSIONS

In this paper, we strive to explore an objective data-driven approach for factor investigation on recommendation serendipity, which can be taken as an important attempt in addition to user surveys. We thoroughly researched the literature to extract all possible factors and propose two principles of meaning coverage and factor independence to clarify them. Then we propose a general framework to evaluate each factor's impact on serendipity. Next, we provide one possible implementation approach of the framework, and we conduct comprehensive experiments on large-scale cross-domain datasets, which evaluate the relative importance of different factors in different domains. We also propose a quantifying method of the degree of disentanglement to measure the distance between any two factors, which provides some insight into understanding factor correlation and combining proper factors to improve serendipity. Finally, we observe that the domain features also matter with factor importance, which reflects the main characteristics of user-item interactions in different domains. This can be used to guide serendipity recommendations in more domains.

It should be noted that the patterns or results identified in this study can serve as an example to explore the impact of serendipity and its corresponding factors. The findings may not be generalized to all domains or applications. However, the proposed method is general and can be applied in other real-life scenarios, and the disentangled seven factors can serve as a basis for constructing serendipity in different domains. There are several directions for future work: (1) It is promising to explore the seven factors to enhance serendipity recommendation in various domains, particularly in those where serendipity recommendations have not been fully studied, e.g., smart education, lifelong learning, science popularization, and so on. (2) More implementations can be tried on factor disentanglement and the impact evaluation framework. (3) It will be interesting to design interactive tools that allow users to select their serendipity factors [10]. (4) The research method can also be exploited to deal with other complex and ambiguous concepts.

REFERENCES

[1] Fakhri Abbas. 2018. Serendipity in Recommender System: A Holistic Overview. In *AICCSA 2018*. IEEE Computer Society, 1–2. https://doi.org/10.1109/AICCSA.2018.8612895

[2] Fakhri Abbas and Xi Niu. 2019. Computational Serendipitous Recommender System Frameworks: A Literature Survey. In *AICCSA 2019*. IEEE Computer Society, 1–8. https://doi.org/10.1109/AICCSA47632.2019.9035339

[3] Panagiotis Adamopoulos and Alexander Tuzhilin. 2014. On Unexpectedness in Recommender Systems: Or How to Better Expect the Unexpected. *ACM Trans. Intell. Syst. Technol.* 5, 4 (2014), 54:1–54:32. https://doi.org/10.1145/2559952

[4] Takayuki Akiyama, Kiyohiro Obara, and Masaaki Tanizaki. 2010. Proposal and Evaluation of Serendipitous Recommendation Method Using General Unexpectedness. In *PRSAT 2010*, Vol. 676. CEUR-WS.org, 3–10. http://ceur-ws.org/Vol-676/paper1.pdf

[5] 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 Conference on Creativity & Cognition, 2009*. ACM, 305–314. https://doi.org/10.1145/1640233.1640279

[6] Taushif Anwar and V. Uma. 2019. A Review of Recommender System and Related Dimensions. *Data, Engineering and Applications* (2019), 3–10. https://doi.org/10.1007/978-981-13-6347-4_1

[7] Miriam El Khoury Badran, Jacques Bou Abdo, Wissam Al Jurdi, and Jacques Demerjian. 2019. Adaptive Serendipity for Recommender Systems: Let It Find You. In *ICAART 2019*. SciTePress, 739–745. https://doi.org/10.5220/0007409507390745

[8] Giacomo Balloccu, Ludovico Boratto, Gianni Fenu, and Mirko Marras. 2022. Post Processing Recommender Systems with Knowledge Graphs for Recency, Popularity, and Diversity of Explanations. In *Proc ACM SIGIR 2022* (Madrid, Spain). ACM, New York, NY, USA, 646–656. https://doi.org/10.1145/3477495.3532041

[9] Pia Borlund. 2003. The Concept of Relevance in IR. *J. Assoc. Inf. Sci. Technol.* 54, 10 (2003), 913–925. https://doi.org/10.1002/asi.10286

[10] Jacquelyn Burkell, Anabel Quan-Haase, and Victoria L. Rubin. 2012. Promoting Serendipity Online: Recommendations for Tool Design. In *Proceedings of the 2012 IConference* (Toronto, Ontario, Canada) *(iConference '12)*. Association for Computing Machinery, New York, NY, USA, 525–526. https://doi.org/10.1145/2132176.2132274

[11] Natthanun Chantanurak, Proadpran Punyabukkana, and Atiwong Suchato. 2016. Video Recommender System using textual data: Its application on LMS and Serendipity evaluation. In *TALE 2016*. IEEE, 289–295. https://doi.org/10.1109/TALE.2016.7851809

[12] Jiaju Chen, Wang Wenjie, Chongming Gao, Peng Wu, Jianxiong Wei, and Qingsong Hua. 2024. Treatment Effect Estimation for User Interest Exploration on Recommender Systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information*

Retrieval (Washington DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 1861–1871. https://doi.org/10.1145/3626772.3657736

[13] Li Chen, Yonghua Yang, Ningxia Wang, Keping Yang, and Quan Yuan. 2019. How Serendipity Improves User Satisfaction with Recommendations? A Large-Scale User Evaluation. In *WWW 2019*. ACM, 240–250. https://doi.org/10.1145/3308558.3313469

[14] Laming Chen, Guoxin Zhang, and Eric Zhou. 2018. Fast Greedy MAP Inference for Determinantal Point Process to Improve Recommendation Diversity. In *NeurIPS 2018*. 5627–5638.

[15] Yifan Chen, Pengjie Ren, Yang Wang, and Maarten de Rijke. 2019. Bayesian Personalized Feature Interaction Selection for Factorization Machines. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Paris, France) *(SIGIR'19)*. Association for Computing Machinery, New York, NY, USA, 665–674. https://doi.org/10.1145/3331184.3331196

[16] Yifan Chen, Yang Wang, Pengjie Ren, Meng Wang, and Maarten de Rijke. 2022. Bayesian feature interaction selection for factorization machines. *Artificial Intelligence* 302 (2022), 103589. https://doi.org/10.1016/j.artint.2021.103589

[17] Anup Anand Deshmukh, Pratheeksha Nair, and Shrisha Rao. 2018. A Scalable Clustering Algorithm for Serendipity in Recommender Systems. In *ICDM Workshops 2018*. IEEE, 1279–1288. https://doi.org/10.1109/ICDMW.2018.00182

[18] Yuntao Du, Xinjun Zhu, Lu Chen, Baihua Zheng, and Yunjun Gao. 2022. HAKG: Hierarchy-Aware Knowledge Gated Network for Recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain) *(SIGIR '22)*. ACM, New York, NY, USA, 1390–1400. https://doi.org/10.1145/3477495.3531987

[19] Xiangyu Fan and Xi Niu. 2018. Implementing and Evaluating Serendipity in Delivering Personalized Health Information. *ACM Trans. Manag. Inf. Syst.* 9, 2 (2018), 7:1–7:19. https://doi.org/10.1145/3205849

[20] Ziwei Fan, Ke Xu, Zhang Dong, Hao Peng, Jiawei Zhang, and Philip S. Yu. 2023. Graph Collaborative Signals Denoising and Augmentation for Recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, Hsin-Hsi Chen, Wei-Jou (Edward) Duh, Hen-Hsen Huang, Makoto P. Kato, Josiane Mothe, and Barbara Poblete (Eds.). ACM, 2037–2041. https://doi.org/10.1145/3539618.3591994

[21] Andres Ferraro. 2019. Music Cold-start and Long-tail Recommendation: Bias in Deep Representations. In *RecSys 2019*. ACM, 586–590.

[22] Zhe Fu, Xi Niu, and Mary Lou Maher. 2024. Deep Learning Models for Serendipity Recommendations: A Survey and New Perspectives. *ACM Comput. Surv.* 56, 1 (2024), 19:1–19:26. https://doi.org/10.1145/3605145

[23] Zhe Fu, Xi Niu, and Li Yu. 2023. Wisdom of Crowds and Fine-Grained Learning for Serendipity Recommendations. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 739–748.

[24] Lu Gan, Diana Nurbakova, Léa Laporte, and Sylvie Calabretto. 2020. Enhancing Recommendation Diversity using Determinantal Point Processes on Knowledge Graphs. In *SIGIR 2020*. ACM, 2001–2004. https://doi.org/10.1145/3397271.3401213

[25] Chongming Gao, Shiqi Wang, Shijun Li, Jiawei Chen, Xiangnan He, Wenqiang Lei, Biao Li, Yuan Zhang, and Peng Jiang. 2024. CIRS: Bursting Filter Bubbles by Counterfactual Interactive Recommender System. *ACM Trans. Inf. Syst.* 42, 1 (2024), 14:1–14:27. https://doi.org/10.1145/3594871

[26] Mouzhi Ge, Carla Delgado-Battenfeld, and Dietmar Jannach. 2010. Beyond Accuracy: Evaluating Recommender Systems by Coverage and Serendipity. In *RecSys 2010*. ACM, 257–260. https://doi.org/10.1145/1864708.1864761

[27] Xiaoyu Ge, Panos K. Chrysanthis, Konstantinos Pelechrinis, Demetrios Zeinalipour-Yazti, and Mohamed A. Sharaf. 2020. Serendipity-based Points-of-Interest Navigation. *ACM Trans. Internet Techn.* 20, 4 (2020), 33:1–33:32. https://doi.org/10.1145/3391197

[28] Xiaoyu Ge, Ameya Daphalapurkar, Manali Shimpi, Darpun Kohli, Konstantinos Pelechrinis, Panos K. Chrysanthis, and Demetrios Zeinalipour-Yazti. 2017. Data-Driven Serendipity Navigation in Urban Places. In *ICDCS 2017*. 2501–2504. https://doi.org/10.1109/ICDCS.2017.286

[29] Marco de Gemmis, Pasquale Lops, and Giovanni Semeraro. 2017. Emotion Detection Techniques for the Evaluation of Serendipitous Recommendations. In *Emotions and Personality in Personalized Services - Models, Evaluation and Applications*. Springer, 357–376. https://doi.org/10.1007/978-3-319-31413-6_17

[30] Marco de Gemmis, Pasquale Lops, Giovanni Semeraro, and Cataldo Musto. 2015. An Investigation on The Serendipity Problem in Recommender Systems. *Inf. Process. Manag.* 51, 5 (2015), 695–717. https://doi.org/10.1016/j.ipm.2015.06.008

[31] Marco Gori and Augusto Pucci. 2007. ItemRank: A Random-Walk Based Scoring Algorithm for Recommender Engines. In *IJCAI 2007*. 2766–2771. http://ijcai.org/Proceedings/07/Papers/444.pdf

[32] Jiqing Gu, Chao Song, Wenjun Jiang, Xiaomin Wang, and Ming Liu. 2020. Enhancing Personalized Trip Recommendation with Attractive Routes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34(01). 662–669. https://doi.org/10.1609/aaai.v34i01.5407

[33] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, YongDong Zhang, and Meng Wang. 2020. LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation. In *SIGIR 2020*. ACM, 639–648. https://doi.org/10.1145/3397271.3401063

[34] Jizhou Huang, Shiqiang Ding, Haifeng Wang, and Ting Liu. 2018. Learning to Recommend Related Entities With Serendipity for Web Search Users. *ACM Trans. Asian Low Resour. Lang. Inf. Process.* 17, 3 (2018), 25:1–25:22. https://doi.org/10.1145/3185663

[35] Leo Iaquinta, Marco de Gemmis, Pasquale Lops, Giovanni Semeraro, Michele Filannino, and Piero Molino. 2008. Introducing Serendipity in a Content-Based Recommender System. In *HIS 2008*. IEEE Computer Society, 168–173. https://doi.org/10.1109/HIS.2008.25

[36] Satoko Inoue and Masataka Tokumaru. 2020. Serendipity Recommender System for Academic Disciplines. In *SCIS/ISIS 2020*. IEEE, 1–4. https://doi.org/10.1109/SCISISIS50064.2020.9322747

[37] Yitong Ji, Aixin Sun, Jie Zhang, and Chenliang Li. 2023. A Critical Study on Data Leakage in Recommender System Offline Evaluation. *ACM Trans. Inf. Syst.* 41, 3, Article 75 (feb 2023), 27 pages. https://doi.org/10.1145/3569930

[38] Marius Kaminskas and Derek Bridge. 2017. Diversity, Serendipity, Novelty, and Coverage: A Survey and Empirical Analysis of Beyond-Accuracy Objectives in Recommender Systems. *ACM Trans. Interact. Intell. Syst.* 7, 1 (2017), 2:1–2:42. https://doi.org/10.1145/2926720

[39] Komal Kapoor, Vikas Kumar, Loren G. Terveen, Joseph A. Konstan, and Paul R. Schrater. 2015. "I Like to Explore Sometimes": Adapting to Dynamic User Novelty Preferences. In *RecSys 2015*. ACM, 19–26. https://doi.org/10.1145/2792838.2800172

[40] Aleksandra Karpus, Iacopo Vagliano, and Krzysztof Goczyla. 2017. Serendipitous Recommendations Through Ontology-Based Contextual Pre-filtering. In *BDAS 2017*, Vol. 716. 246–259. https://doi.org/10.1007/978-3-319-58274-0_21

[41] Samira Khoshahval, Mahdi Farnaghi, Mohammad Taleai, and Ali Mansourian. 2018. A Personalized Location-Based and Serendipity-Oriented Point of Interest Recommender Assistant Based on Behavioral Patterns. In *Geospatial Technologies for All - Selected Papers of the 21st AGILE Conference on Geographic Information Science*. Springer, 271–289. https://doi.org/10.1007/978-3-319-78208-9_14

[42] Denis Kotkov, Joseph A. Konstan, Qian Zhao, and Jari Veijalainen. 2018. Investigating Serendipity in Recommender Systems Based on Real User Feedback. In *SAC 2018*. ACM, 1341–1350. https://doi.org/10.1145/3167132.3167276

[43] Denis Kotkov, Alan Medlar, Triin Kask, and Dorota Glowacka. 2024. The Dark Matter of Serendipity in Recommender Systems. In *Proceedings of the 2024 Conference on Human Information Interaction and Retrieval* (, Sheffield, United Kingdom,) *(CHIIR '24)*. ACM, New York, NY, USA, 108–118. https://doi.org/10.1145/3627508.3638342

[44] Denis Kotkov, Jari Veijalainen, and Shuaiqiang Wang. 2020. How Does Serendipity Affect Diversity in Recommender Systems? A Serendipity-oriented Greedy Algorithm. *Computing* 102, 2 (2020), 393–411. https://doi.org/10.1007/s00607-018-0687-5

[45] Denis Kotkov, Shuaiqiang Wang, and Jari Veijalainen. 2016. A Survey of Serendipity in Recommender Systems. *Knowl. Based Syst.* 111 (2016), 180–192. https://doi.org/10.1016/j.knosys.2016.08.014

[46] David M. Kroenke. 1992. *Database Processing (4th Ed.): Fundamentals, Design Implementation*. Macmillan Publishing Co., Inc., USA.

[47] Hyokmin Kwon, Jaeho Han, and Kyungsik Han. 2020. ART (Attractive Recommendation Tailor): How the Diversity of Product Recommendations Affects Customer Purchase Preference in Fashion Industry?. In *CIKM 2020*. ACM, 2573–2580. https://doi.org/10.1145/3340531.3412687

[48] Neal Lathia, Stephen Hailes, Licia Capra, and Xavier Amatriain. 2010. Temporal Diversity in Recommender Systems. In *SIGIR 2010*. ACM, 210–217. https://doi.org/10.1145/1835449.1835486

[49] John Kalung Leung, Igor Griva, and William G. Kennedy. 2020. Text-based Emotion Aware Recommender. *International Journal on Natural Language Computing* (2020), 101–114. https://doi.org/10.5121/csit.2020.101009

[50] Pan Li, Maofei Que, Zhichao Jiang, YAO HU, and Alexander Tuzhilin. 2020. PURS: Personalized Unexpected Recommender System for Improving User Satisfaction. In *Proceedings of the 14th ACM Conference on Recommender Systems* (Virtual Event, Brazil) *(RecSys '20)*. ACM, New York, NY, USA, 279–288. https://doi.org/10.1145/3383313.3412238

[51] Xueqi Li, Wenjun Jiang, Weiguang Chen, Jie Wu, and Guojun Wang. 2019. HAES: A New Hybrid Approach for Movie Recommendation with Elastic Serendipity. In *CIKM 2019*. ACM, 1503–1512. https://doi.org/10.1145/3357384.3357868

[52] Xueqi Li, Wenjun Jiang, Weiguang Chen, Jie Wu, Guojun Wang, and Kenli Li. 2020. Directional and Explainable Serendipity Recommendation. In *WWW 2020*. ACM / IW3C2, 122–132. https://doi.org/10.1145/3366423.3380100

[53] Jianxun Lian, Iyad Batal, Zheng Liu, Akshay Soni, Eun Yong Kang, Yajun Wang, and Xing Xie. 2021. Multi-Interest-Aware User Modeling for Large-Scale Sequential Recommendations. *CoRR* abs/2102.09211 (2021). arXiv:2102.09211 https://arxiv.org/abs/2102.09211

[54] Zihan Lin, Hui Wang, Jingshu Mao, Wayne Xin Zhao, Cheng Wang, Peng Jiang, and Ji-Rong Wen. 2022. Feature-aware Diversified Re-ranking with Disentangled Representations for Relevant Recommendation. In *Proc. SIGKDD 2022*. ACM. https://doi.org/10.1145/3534678.3539130

[55] Pasquale Lops, Marco de Gemmis, and Giovanni Semeraro. 2011. Content-based Recommender Systems: State of the Art and Trends. In *Recommender Systems Handbook*. Springer, 73–105. https://doi.org/10.1007/978-0-387-85820-3_3

[56] Qiuxia Lu, Tianqi Chen, Weinan Zhang, Diyi Yang, and Yong Yu. 2012. Serendipitous Personalized Ranking for Top-N Recommendation. In *Web Intelligence 2012*. IEEE Computer Society, 258–265. https://doi.org/10.1109/WI-IAT.2012.135

[57] Valentina Maccatrozzo, Manon Terstall, Lora Aroyo, and Guus Schreiber. 2017. SIRUP: Serendipity In Recommendations via User Perceptions. In *IUI 2017*. ACM, 35–44. https://doi.org/10.1145/3025171.3025185

[58] Maake Benard Magara, Sunday O. Ojo, and Tranos Zuva. 2018. Towards a Serendipitous Research Paper Recommender System Using Bisociative Information Networks (BisoNets). In *2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD)*. 1–6. https://doi.org/10.1109/ICABCD.2018.8465475

[59] Andrii Maksai, Florent Garcin, and Boi Faltings. 2015. Predicting Online Performance of News Recommender Systems Through Richer Evaluation Metrics. In *RecSys 2015*. ACM, 179–186. https://doi.org/10.1145/2792838.2800184

[60] Lori McCay-Peet and Elaine G Toms. 2013. Proposed Facets of A Serendipitous Digital Environment. (2013), 688–691. https://api.semanticscholar.org/CorpusID:111911140

[61] Lori McCay-Peet and Elaine G. Toms. 2015. Investigating Serendipity: How It Unfolds and What May Influence It. *Journal of the Association for Information Science and Technology* 66, 7 (2015), 1463–1476. https://doi.org/10.1002/asi.23273 arXiv:https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/asi.23273

[62] Alan Menk, Laura Sebastia, and Rebeca Ferreira. 2017. CURUMIM: A Serendipitous Recommender System for Tourism Based on Human Curiosity. In *2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI)*. 788–795. https://doi.org/10.1109/ICTAI.2017.00124

[63] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, Yoshua Bengio and Yann

LeCun (Eds.). http://arxiv.org/abs/1301.3781

[64] Barbara Minto. 1996. The Minto Pyramid Principle: Logic in Writing, Thinking and Problem Solving. *Minto International* (1996).

[65] Tomoko Murakami, Koichiro Mori, and Ryohei Orihara. 2007. Metrics for Evaluating the Serendipity of Recommendation Lists. In *JSAI 2007*, Vol. 4914. Springer, 40–46. https://doi.org/10.1007/978-3-540-78197-4_5

[66] Chifumi Nishioka, Jörn Hauke, and Ansgar Scherp. 2020. Influence of Tweets and Diversification on Serendipitous Research Paper Recommender Systems. *PeerJ Comput. Sci.* 6 (2020), e273. https://doi.org/10.7717/peerj-cs.273

[67] Xi Niu, Fakhri Abbas, Mary Lou Maher, and Kazjon Grace. 2018. Surprise Me If You Can: Serendipity in Health Information. In *CHI 2018*. ACM, 1–12. https://doi.org/10.1145/3173574.3173597

[68] Arseto Satriyo Nugroho, Igi Ardiyanto, and Teguh Bharata Adji. 2021. User Curiosity Factor in Determining Serendipity of Recommender System. *International Journal of Information Technology and Electrical Engineering* 5, 3 (2021), 75–81. https://doi.org/10.22146/ijitee.67553

[69] Kenta Oku and Fumio Hattori. 2012. User Evaluation of Fusion-based Recommender Systems for Serendipity-oriented Recommendation. In *RUE 2012*, Vol. 910. CEUR-WS.org, 39–44. http://ceur-ws.org/Vol-910/paper9.pdf

[70] Zachary A. Pardos and Weijie Jiang. 2020. Designing for serendipity in a university course recommendation system. In *Learning Analytics and Knowledge 2020*. ACM, 350–359. https://doi.org/10.1145/3375462.3375524

[71] DaEun Park, Jongmo Kim, and Mye M. Sohn. 2019. Serendipity-Based Recommendation Framework for SNS Users Using Tie Strength and Relation Clustering. In *IMIS 2019*, Vol. 994. Springer, 636–645. https://doi.org/10.1007/978-3-030-22263-5_60

[72] Daniil Pastukhov, Stanislav Kuznetsov, Vojtěch Vančura, and Pavel Kordík. 2022. Offline Evaluation of the Serendipity in Recommendation Systems. In *2022 IEEE 17th International Conference on Computer Sciences and Information Technologies (CSIT)*. 597–601. https://doi.org/10.1109/CSIT56902.2022.10000782

[73] Xiangjun Peng, Hongzhi Zhang, Xiaosong Zhou, Shuolei Wang, Xu Sun, and Qingfeng Wang. 2020. CHESTNUT: Improve Serendipity in Movie Recommendation by an Information Theory-Based Collaborative Filtering Approach. In *HCII 2020*, Vol. 12185. Springer, 78–95. https://doi.org/10.1007/978-3-030-50017-7_6

[74] Tieyun Qian, Yile Liang, Qing Li, Xuan Ma, Ke Sun, and Zhiyong Peng. 2022. Intent Disentanglement and Feature Self-supervision for Novel Recommendation. *IEEE Transactions on Knowledge and Data Engineering* (2022), 1–14. https://doi.org/10.1109/TKDE.2022.3175536

[75] Urbano Reviglio. 2019. Serendipity As An Emerging Design Principle of the Infosphere: Challenges and Opportunities. *Ethics and Information Technology* 21, 2 (2019), 151–166. Publisher: Springer.

[76] Nur Izyan Yasmin Saat, Shahrul Azman Mohd Noah, and Masnizah Mohd. 2018. Towards Serendipity for Content–Based Recommender Systems. *International Journal on Advanced Science, Engineering and Information Technology* 8 (2018), 1763–1769.

[77] Javier Sanz-Cruzado and Pablo Castells. 2018. Enhancing Structural Diversity in Social Networks by Recommending Weak Ties. In *RecSys 2018*. ACM, 233–241. https://doi.org/10.1145/3240323.3240371

[78] I. Schmitt and S. Conrad. 1998. Re-Engineering Object-Oriented Database Schemata by Concept Analysis. In *Preproc. of the 7th International Workshop on Foundations of Models and Languages for Data and Objects (FoMLaDO '98), Timmel, Deutschland, Oktober 1998*. Fachbereich Informatik, Universität Dortmund, 190–198.

[79] Guy Shani and Asela Gunawardana. 2011. Evaluating Recommendation Systems. In *Recommender Systems Handbook*. Springer, 257–297. https://doi.org/10.1007/978-0-387-85820-3_8

[80] Zihua Si, Zhongxiang Sun, Xiao Zhang, Jun Xu, Yang Song, Xiaoxue Zang, and Ji-Rong Wen. 2023. Enhancing Recommendation with Search Data in a Causal Learning Manner. *ACM Trans. Inf. Syst.* 41, 4, Article 111 (apr 2023), 31 pages. https://doi.org/10.1145/3582425

[81] Andrei Martins Silva, Fernando Henrique da Silva Costa, Alexandra Katiuska Ramos Diaz, and Sarajane Marques Peres. 2018. Exploring Coclustering for Serendipity Improvement in Content-Based Recommendation. In *IDEAL 2018*, Vol. 11314. Springer, 317–327. https://doi.org/10.1007/978-3-030-03493-1_34

[82] Haoran Tang, Shiqing Wu, Guandong Xu, and Qing Li. 2023. Dynamic Graph Evolution Learning for Recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (, Taipei, Taiwan,) *(SIGIR '23)*. Association for Computing Machinery, New York, NY, USA, 1589–1598. https://doi.org/10.1145/3539618.3591674

[83] Jan E Trost. 1986. Statistically Nonrepresentative Stratified Sampling: A Sampling Technique for Qualitative Studies. *Qualitative sociology* 9, 1 (1986), 54–57. Publisher: Springer.

[84] Noa Tuval. 2019. Exploring the Potential of the Resolving Sets Model for Introducing Serendipity to Recommender Systems. In *UMAP 2019*. ACM, 353–356. https://doi.org/10.1145/3320435.3323467

[85] Liangtian Wan, Yuyuan Yuan, Feng Xia, and Huan Liu. 2021. To your surprise: Identifying serendipitous collaborators. *IEEE Transactions on Big Data* 7, 3 (2021), 574–589. https://doi.org/10.1109/TBDATA.2019.2921567

[86] Changdong Wang, Zhihong Deng, Jianhuang Lai, and Philip S. Yu. 2019. Serendipitous Recommendation in E-Commerce Using Innovator-Based Collaborative Filtering. *IEEE Trans. Cybern.* 49, 7 (2019), 2678–2692. https://doi.org/10.1109/TCYB.2018.2841924

[87] Ningxia Wang, Li Chen, and Yonghua Yang. 2020. The Impacts of Item Features and User Characteristics on Users' Perceived Serendipity of Recommendations. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization* (Genoa, Italy) *(UMAP '20)*. 266–274. https://doi.org/10.1145/3340631.3394863

[88] Wenjie Wang, Fuli Feng, Liqiang Nie, and Tat-Seng Chua. 2022. User-Controllable Recommendation Against Filter Bubbles. In *Proc. ACM SIGIR* (Madrid, Spain) *(SIGIR '22)*. 1251–1261. https://doi.org/10.1145/3477495.3532075

[89] Yifan Wang, Weizhi Ma, Min Zhang, Yiqun Liu, and Shaoping Ma. 2023. A Survey on the Fairness of Recommender Systems. *ACM Trans. Inf. Syst.* 41, 3 (feb 2023), Article No: 52, pages:1–43. https://doi.org/10.1145/3547333

[90] Zongyi Wang, Yanyan Zou, Anyu Dai, Linfang Hou, Nan Qiao, Luobao Zou, Mian Ma, Zhuoye Ding, and Sulong Xu. 2023. An Industrial Framework for Personalized Serendipitous Recommendation in E-commerce. In *Proceedings of the 17th ACM Conference on Recommender Systems* (Singapore, Singapore) *(RecSys '23)*. Association for Computing Machinery, New York, NY, USA, 1015–1018. https://doi.org/10.1145/3604915.3610234

[91] Yuanbo Xu, Yongjian Yang, En Wang, Jiayu Han, Fuzhen Zhuang, Zhiwen Yu, and Hui Xiong. 2020. Neural Serendipity Recommendation: Exploring the Balance between Accuracy and Novelty with Sparse Explicit Feedback. *ACM Trans. Knowl. Discov. Data* 14, 4 (2020), 50:1–50:25. https://doi.org/10.1145/3396607

[92] Zhenzhen Xu, Yuyuan Yuan, Haoran Wei, and Liangtian Wan. 2019. A Serendipity-biased Deepwalk for Collaborators Recommendation. *PeerJ Comput. Sci.* 5 (2019), e178. https://doi.org/10.7717/peerj-cs.178

[93] Yongjian Yang, Yuanbo Xu, En Wang, Jiayu Han, and Zhiwen Yu. 2018. Improving Existing Collaborative Filtering Recommendations via Serendipity-Based Algorithm. *IEEE Transactions on Multimedia* 20, 7 (2018), 1888–1900. https://doi.org/10.1109/TMM.2017.2779043

[94] Huan Yu, Ying Wang, Yaning Fan, Sachula Meng, and Rui Huang. 2017. Accuracy is Not Enough: Serendipity Should Be Considered More. In *International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing*. Springer, 231–241.

[95] Chengyuan Zhang, Yang Wang, Lei Zhu, Jiayu Song, and Hongzhi Yin. 2021. Multi-Graph Heterogeneous Interaction Fusion for Social Recommendation. *ACM Trans. Inf. Syst.* 40, 2, Article 28 (Sept. 2021), 26 pages. https://doi.org/10.1145/3466641

[96] Mingwei Zhang, Yang Yang, Rizwan Abbas, Ke Deng, Jianxin Li, and Bin Zhang. 2021. SNPR: A Serendipity-Oriented Next POI Recommendation Model. In *CIKM 2021*. ACM, 2568–2577. https://doi.org/10.1007/s00779-020-01371-w

[97] Xiaokun Zhang, Bo Xu, Youlin Wu, Yuan Zhong, Hongfei Lin, and Fenglong Ma. 2024. FineRec: Exploring Fine-grained Sequential Recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 1599–1608. https://doi.org/10.1145/3626772.3657761

[98] Xiaokun Zhang, Bo Xu, Liang Yang, Chenliang Li, Fenglong Ma, Haifeng Liu, and Hongfei Lin. 2022. Price DOES Matter! Modeling Price and Interest Preferences in Session-Based Recommendation. In *Proc. ACM SIGIR* (Madrid, Spain) *(SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 1684–1693. https://doi.org/10.1145/3477495.3532043

[99] Gang Zhao, Mong-Li Lee, Wynne Hsu, and Wei Chen. 2012. Increasing Temporal Diversity with Purchase Intervals. In *SIGIR 2012*. ACM, 165–174. https://doi.org/10.1145/2348283.2348309

[100] Xing Zhao, Ziwei Zhu, and James Caverlee. 2021. Rabbit Holes and Taste Distortion: Distribution-Aware Recommendation with Evolving Interests. In *Proceedings of the Web Conference 2021* (Ljubljana, Slovenia) *(WWW '21)*. Association for Computing Machinery, New York, NY, USA, 888–899. https://doi.org/10.1145/3442381.3450099

[101] Qianru Zheng, Chi-Kong Chan, and Horace H. S. Ip. 2015. An Unexpectedness-Augmented Utility Model for Making Serendipitous Recommendation. In *ICDM 2015*, Vol. 9165. Springer, 216–230. https://doi.org/10.1007/978-3-319-20910-4_16

[102] Yu Zheng, Chen Gao, Jianxin Chang, Yanan Niu, Yang Song, Depeng Jin, and Yong Li. 2022. Disentangling Long and Short-Term Interests for Recommendation. In *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, Frédérique Laforest, Raphaël Troncy, Elena Simperl, Deepak Agarwal, Aristides Gionis, Ivan Herman, and Lionel Médini (Eds.). ACM, 2256–2267. https://doi.org/10.1145/3485447.3512098

[103] Yu Zheng, Chen Gao, Liang Chen, Depeng Jin, and Yong Li. 2021. DGCN: Diversified Recommendation with Graph Convolutional Networks. In *Proceedings of the Web Conference 2021* (Ljubljana, Slovenia) *(WWW '21)*. Association for Computing Machinery, New York, NY, USA, 401–412. https://doi.org/10.1145/3442381.3449835

[104] Xiaosong Zhou, Zhan Xu, Xu Sun, and Qingfeng Wang. 2017. A New Information Theory-based Serendipitous Algorithm Design. In *International Conference on Human Interface and the Management of Information*. Springer, 314–327.

[105] Xinjun Zhu, Yuntao Du, Yuren Mao, Lu Chen, Yujia Hu, and Yunjun Gao. 2023. Knowledge-refined Denoising Network for Robust Recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (, Taipei, Taiwan,) *(SIGIR '23)*. Association for Computing Machinery, New York, NY, USA, 362–371. https://doi.org/10.1145/3539618.3591707

[106] Yaochen Zhu, Liang Wu, Qi Guo, Liangjie Hong, and Jundong Li. 2024. Collaborative Large Language Model for Recommender Systems. In *Proceedings of the ACM Web Conference 2024* (Singapore, Singapore) *(WWW '24)*. Association for Computing Machinery, New York, NY, USA, 3162–3172. https://doi.org/10.1145/3589334.3645347

[107] Reza Jafari Ziarani and Reza Ravanmehr. 2021. Deep neural network approach for a serendipity-oriented recommendation system. *Expert Syst. Appl.* 185 (2021), 115660. https://doi.org/10.1016/j.eswa.2021.115660

[108] Reza Jafari Ziarani and Reza Ravanmehr. 2021. Serendipity in Recommender Systems: A Systematic Literature Review. *Journal of Computer Science and Technology* 36, 2 (2021), 375–396.

A APPENDIX

A.1 Frequently Mentioned Serendipity Factors in Literature

We list some representative and frequently mentioned serendipity factors in the literature, shown in Table 5. It can be seen that different works exploit different factors to construct recommendation serendipity; moreover, in various works, the same factor name may express different meanings and vice versa. The ambiguity and inconsistency between the factors' names and meanings significantly impact the understanding of recommendation serendipity and hinder the follow-up research. Hence, there is a strong necessity to clarify those factors.

Table 5. Frequently Mentioned Serendipity Factors in the Literature.

Factors	Source	Statement
unexpectedness and relevance	[81]	**unexpectedness**: the number of ratings given to i is smaller than the average number of ratings given to all items; **relevance**: the rating given to i is bigger than the average rating by u.
	[7]	**unexpectedness**: between a lower limit and an upper limit on the distance of recommended items from expectations
	[66]	**unexpectedness**: inferred by a primitive strategy
	[70]	user perceived unexpectedness of result with successfulness
unexpectedness, relevance and novelty	[45]	**unexpectedness**: significantly differ from user profile; **relevance**: a user likes, consumes or is interested in; **novelty**: a user has never consumed the item.
	[42]	**novelty**: in three ways, new to system, new to user and forgotten items [39].
	[76]	**unexpectedness**: a positive response from users; **relevance**: some similarities with user profile; **novelty**: new items to users.
unexpectedness, relevance, novelty, and timeliness	[13]	**novelty**: from a category/domain outside of the user's profile
unexpectedness, relevance, and value	[92]	**unexpectedness**: have different research topics from the target scholars; **relevance**: the proximity between two nodes in co-author network; **value**: the influence of this collaborator.
unexpectedness, relevance and positive surprise	[27]	**unexpectedness**: achieved by randomness, and non-determinism algorithms.
unexpectedness, relevance and diversity	[17]	**unexpectedness**: the difference between recommendations made before and after considering serendipity; **relevance**: not completely dissimilar from a user's interests; **diversity**: the differences among recommended items.
unexpectedness and usefulness	[6]	not appear to be connected and will become interesting
	[59]	**unexpectedness** penalizes the most popular items; **usefulness**: accuracy
	[30]	**unexpectedness**: the items in the long tail
	[26]	**unexpectedness**: based on a primitive model; **usefulness**: judged by users
unexpectedness, usefulness and insight	[73]	**unexpectedness**: based on a primitive prediction model; **usefulness**: the potential value; **insight**: "making connections" to a target user's profile. [104]
unexpectedness, novelty, and value	[1]	**value**: interestingness [5], usefulness [3]; **novelty**: unknown and different item, where unknown items include items that the user never consumed before and different items include items that are different from the user's profile [38, 45].
unexpectedness and value	[67]	recommending items that are valuable to the user, but do not contain content that the user was expecting.
unexpectedness, value and insight	[104]	**unexpectedness**: the encountered information should be unexpected or a surprise to the information actor; **value**: useful and beneficial to the information actor; **insight**: an ability to find some clue in the current environment, then "making connections" with one's previous knowledge or experience.
unexpectedness and attractiveness	[29]	**unexpectedness**: the deviation from a benchmark model [31, 65]; **attractiveness**: determined in terms of closeness to the user profile [55]
unexpectedness, interestingness and relatedness	[34]	**unexpectedness**: the relation between the entity and the query should not have been otherwise discovered by the user; **interestingness**: the entity should engage the interest of the user when searching for the query; **relatedness**: relevant to the query that a user is searching for.
unexpectedness and accuracy	[41]	**unexpected**: dissimilar to user profile or historical items
unexpectedness and interest	[19]	**unexpectedness**: three ways are employed to recommend items with different distances from user's interests
unexpectedness, good surprise and preference	[11]	**unexpectedness**: determined by users
unexpectedness and high quality	[94]	**unexpectedness**: distant from users' latest preferences
value and surprise	[67]	**value**: willing to pay and experienced utility. Expectation for an information object is based on the expected likelihood of a user seeing such an information object; **surprise**: a violation of such expectation.
surprise and success	[79]	**surprise**: far from the user profile; **success**: similar to user's history
accuracy and difference	[52]	**accuracy**: similarity to users' long-term preferences and short-term demands and recommendation accuracy; **difference**: difference from users' history and diversity among recommendations
genre accuracy and content difference	[51]	accuracy in genre-level preference and difference in collaborative filtering
high satisfaction and low interest	[91]	**satisfaction**: items with high ratings; **interest**: items with ratings