

Result Diversification in Search and Recommendation: A Survey

Haolun Wu[†], Yansen Zhang[†], Chen Ma*, Fuyuan Lyu, Bowei He, Bhaskar Mitra, and Xue Liu *Fellow, IEEE*

Abstract—Diversifying return results is an important research topic in retrieval systems in order to satisfy both the various interests of customers and the equal market exposure of providers. There has been growing attention on diversity-aware research during recent years, accompanied by a proliferation of literature on methods to promote diversity in search and recommendation. However, diversity-aware studies in retrieval systems lack a systematic organization and are rather fragmented. In this survey, we are the first to propose a unified taxonomy for classifying the metrics and approaches of diversification in both search and recommendation, which are two of the most extensively researched fields of retrieval systems. We begin the survey with a brief discussion of why diversity is important in retrieval systems, followed by a summary of the various diversity concerns in search and recommendation, highlighting their relationship and differences. For the survey’s main body, we present a unified taxonomy of diversification metrics and approaches in retrieval systems, from both the search and recommendation perspectives. In the later part of the survey, we discuss the open research questions of diversity-aware research in search and recommendation in an effort to inspire future innovations and encourage the implementation of diversity in real-world systems. We maintain an implementation for classical diversification metrics and methods summarized in this survey at <https://github.com/Forrest-Stone/Diversity>.

1 INTRODUCTION

WITH the ever-growing volume of online information, users can easily access an increasingly vast number of online products and services. To alleviate information overload and expedite the acquisition of information, information retrieval systems have emerged and begun to play important roles in modern society. Search and recommendation are two of the most important applications of retrieval systems; both can be viewed as ranking systems that output an ordered list. Search systems aim to retrieve relevant entities with respect to the information need(s) behind a query launched by users from a collection of resources. Recommendation systems utilize the user-item interaction history to predict personalized user interests, hence recommending potentially satisfactory items to users.

For a long time, *relevance* dominates the research in both search and recommendation, where the key is to measure if the system is able to retrieve those items that are regarded as “relevant” given part of the ground truth labels [1, 2, 3]. Although these systems are able to retrieve or recommend the most relevant items, they may jeopardize the utilities of stakeholders in the system. Take the recommendation scenario as an example, systems with only high relevance have potential harms for both sides of the two most critical stake-

holders: customers (the user side) and providers (the item side). Customers generally suffer from the *redundancy issue* that recommends redundant or very similar items, which may cause “filter bubble” [4] and harm the customers’ satisfaction in the long term. For instance, a movie recommendation system keeping recommending Marvel action movies to a customer who once clicked “*The Batman*” may block out the opportunity for her from observing other genres of movies. This does not necessarily indicate the customer only likes Marvel action movies. It is highly possible that the customer has various interests but the recommender never explores other choices, thus jeopardizing the customer’s long-term user experience. Providers generally suffer from the *exposure unfairness* due to the “super-star” [5] economy phenomenon where a very small number of most popular items and providers take up an extremely large proportion of exposure to customers. Such unfairness may make those new-coming or less popular providers feel disappointed in attracting customers and finally quit the platforms. Once only several most popular providers remain, monopoly is highly possible, harming a healthy marketplace and society.

From the perspective of search which is less concerned with personalization compared to recommendation, similar limitations occur when merely focusing on relevance. Assuming there is an image search system, it will always retrieve images for “jaguar vehicles” when the query “jaguar” is entered. Although it cannot be disputed that the output of this search system is significantly relevant to the input query, it cannot be considered ideal because the query “jaguar” has additional meanings, such as “jaguar as an animal”, which are never retrieved. Customers can be dissatisfied with the system for the inability to obtain the various desired information. On the other hand, the system may also suffer from the similar *exposure unfairness* described in the recommendation scenario for providers who offer jaguar animal pictures. This is another example

[†]Equal contribution.

*Corresponding author.

- Haolun Wu, Fuyuan Lyu, and Xue Liu are with the School of Computer Science, McGill University, Montreal, Canada. E-mail: haolun.wu@mail.mcgill.ca; fuyuan.lyu@mail.mcgill.ca; xueliu@cs.mcgill.ca
- Yansen Zhang, Chen Ma, and Bowei He are with the Department of Computer Science, City University of Hong Kong, Hong Kong SAR. E-mail: yanshang7-c@my.cityu.edu.hk; chenma@cityu.edu.hk; boweihe2-c@my.cityu.edu.hk
- Bhaskar Mitra is with Microsoft Research, Montreal, Canada. E-mail: bhaskar.mittra@microsoft.com

in the field of search that demonstrates how a singular focus on relevance can have negative effects on numerous stakeholders in a system.

Thus, in recent years, many criteria other than *relevance* have gained tremendous attention in information retrieval systems, and *diversity* is one of the paramount. It has been recognized that diversified search and recommendation can not only increase the likelihood of satisfying various needs of customers in both the short term and long term, but also assist to increase item exposure, especially for those less popular providers [6, 7]. Considering the critical role of diversity in maintaining a satisfactory and healthy information retrieval marketplace, we hereby offer a comprehensive review of definitions, metrics, and techniques of diversity studied in search and recommendation.

Necessity of this Survey. Although many papers have been published on this topic recently, to the best of our knowledge, none of them has provided an unified picture of diversity in both search and recommendation, as well as the corresponding diversity metrics and techniques. We find that the usage of the terminology “diversity” in recent works is usually inconsistent across papers, without a clear claim as to which diversity perspective is emphasized. In addition, some studies lacked an explanation of why they chose particular diversity criteria for measurement. Given the growing awareness of the importance of diversity and the rapid development of diversity techniques in both search and recommendation, we believe our survey can provide a comprehensive summary and organization of the diversity concerns in these fields and offer future researchers a better comprehension of the current state-of-the-art and openness problems on this topic.

Difference with Existing Surveys. A number of surveys in search and recommendation have been published recently, focusing on different perspectives. For instance, in the field of search, Azad and Deepak [8] review the Query Expansion (QE) techniques and Azzopardi [9] summarizes the usage of cognitive bias. In the field of recommendation, Huang et al. [10] provide a summary on privacy protection in recommendation systems and Chen et al. [11] focus on bias and debias techniques. Some other well-cited surveys focus on more general problems in search and recommendation, such as [12] and [13]. However, the perspective of diversity has not been well reviewed in existing search and recommendation surveys. To the best of our knowledge, there exist several surveys on the diversity in recommendation [6, 14, 15], but they do not systematically organize the diversity concerns in both search and recommendation, and the contents are not comprehensive or up-to-date. To offer a comprehensive review of this topic, we make the following contributions in this survey:

- Collecting the latest works and summarizing the types, metrics, and techniques of diversity in both search and recommendation systematically under a unified organization.
- Conducting a detailed analysis and presenting a taxonomy of existing diversity techniques, as well as discussing their strengths and drawbacks.
- Recognizing open research problems and discussing future directions to inspire more research on diversity in search and recommendation.

Papers Collection. We collect over 100 papers that analyze the diversity issues or propose novel techniques in search and recommendation. We first search the related top-tier conferences and journals to find related work, including KDD, NeurIPS, CIKM, RecSys, ICDM, AAAI, WSDM, The WebConf, SIGIR, SIGMOD, TKDD, TKDE, TOIS, etc., with the keywords “search”, “recommendation”, “ranking” or “retrieval” combined with “diversity”, “serendipity” or “coverage” till 2022. We then traverse the citation graph of the identified papers, retaining the papers that focus on diversity. Fig. 1 illustrates the statistics of collected papers with the publication time and venue.

Survey Audience and Organization. This survey is useful for researchers who are new to diversity problems and are looking for a guide to quickly enter this field, as well as those who wish to stay abreast of the most recent diversity strategies in search and recommendation.

The rest of the survey is organized as follows:

- In Section 2 and 3, we summarize the categories and concerns of diversity in search and recommendation.
- In Section 4, we provide the background and preliminaries on search and recommendation systems, followed by listing the notations we used in this survey.
- In Section 5, we review the metrics of diversity generally used in search and recommendation, and systematically categorize them using a unified taxonomy.
- In Section 6 and 7, we review the approaches for enhancing diversity in search and recommendation, from both the offline and online perspectives.
- In Section 8, we discuss the applicability of diversity metrics and approaches to various models.
- In Section 9, we summarize the openness and future directions.

2 DIVERSITY IN SEARCH

Diversifying search results has received attention since the end of last century, where one of the earliest works is Maximal Marginal Relevance (MMR) proposed by Carbonell and Goldstein [16] in 1998. Later, Clarke et al. [17] present a framework that systematically rewards novelty and diversity for measuring information retrieval systems, which promotes a series of works on diversity measurement and improvement in search. As summarized by Radlinski et al. [18] in the 2009 SIGIR Forum, diversity in search can be generally categorized into two classes based on whether the diversity is treated as uncertainty about the information need, or part of the information need. These two concerns are named as (i) *extrinsic diversity* and (ii) *intrinsic diversity* respectively, which are demonstrated as follows.

2.1 Extrinsic Diversity

Extrinsic diversity is related to the situation where uncertainty occurs in search, which can be further divided into the *ambiguity* of the query meaning and the *variability* of the user intent [18]. Generally, these two uncertainties co-occur in a search, as with the query “jaguar”. In other cases, even if there is no *ambiguity* in the query, the user intents may still contain *variability*. For instance, considering a query “BioNTech, Pfizer vaccine”, a patient may seek more information on the vaccination’s effect, whereas a doctor may be

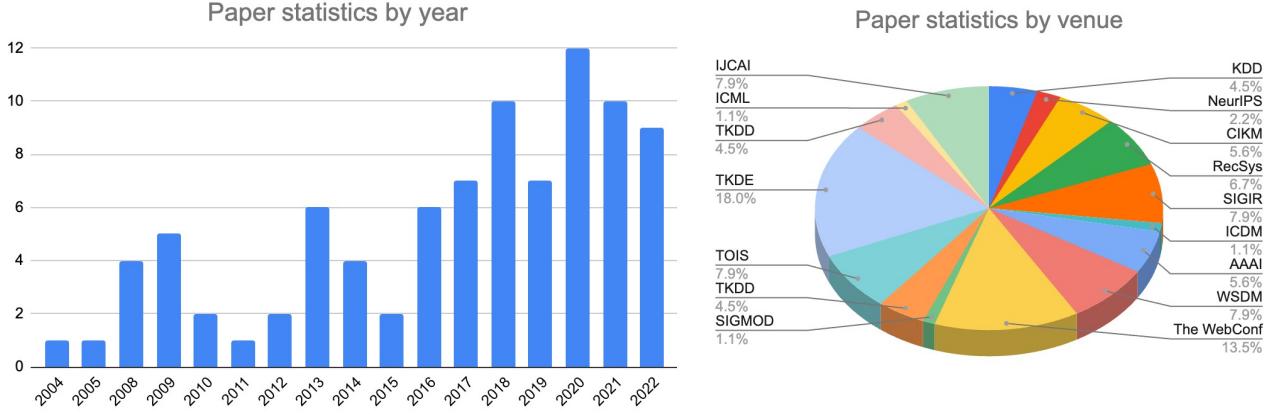


Fig. 1: The statistics of publications related to diversity in search and recommendation with the publication year and venue.

more concerned with the pharmaceutical ingredients and an entrepreneur may be more interested in the company operations of BioNTech. The greater the ability of search results to encompass various query meanings and satisfy multiple user intents, the greater the extrinsic diversity.

2.2 Intrinsic Diversity

Different from extrinsic diversity which treats diversity as uncertainty about the information need, intrinsic diversity treats diversity as part of the information need itself, even given a single well-defined intent. Under this definition, intrinsic diversity can be comprehended as the need for avoiding redundancy in the result lists, which is comparable to the novelty definition by Clarke et al. [17]. The motivation for intrinsic diversity is intuitive: presuming the input query is “jaguar as an animal” with little ambiguity, users may anticipate the search results to contain images of different jaguars from diverse views and angles, rather than the same jaguar with the same view. As such, the less redundancy in the search results, the greater the intrinsic diversity.

To clarify the distinction between extrinsic diversity and intrinsic diversity, the former is a response to a scenario with various search intents, whereas the latter is a response to a scenario with a single search purpose. In real-world cases, both diversity concerns are significant in search and can be measured in a hierarchical and joint way. For instance, a search system may be expected to satisfy various information needs for diverse search intents, while avoiding redundancy for each specific one.

3 DIVERSITY IN RECOMMENDATION

As one of the most significant applications of information retrieval, the diversity in recommendation systems has also been explored. Although the definition of diversity in search is also applicable in the field of recommendation, researchers study the diversity in this field from other perspectives. There are generally two categories of diversity in recommendation across different works: (i) *individual-level diversity* and (ii) *system-level (aggregated) diversity*. Each diversity concern is relevant to one of the two most significant stakeholders: customers and providers, which represent the user side and the item side, respectively. The *individual-level diversity* is relevant to the satisfaction of customers, while the

system-level diversity is relevant to the fairness of providers. In this section, we offer a review and comparison of these two diversity concerns in recommendation systems.

3.1 Individual-level Diversity

The customer is one of the two most significant stakeholders in recommendation systems, whose satisfaction can be influenced by not only the recommendation relevance but also the diversity, serendipity, novelty, etc. Therefore, one category of diversity in the recommendation, individual-level diversity, puts the customer at the core and is intended to quantify the degree to which recommended items are unique to each individual customer. As a result, the recommendation list for each individual is considered separately. From another perspective, individual-level diversity focuses on the problem of how to avoid recommending redundant (but still relevant) items to a customer given the previous recommendation list. Thus, a higher degree of individual-level diversity can provide customers with the opportunity to view more novel items, thereby satisfying diverse demands and facilitating the exploration of various interests.

3.2 System-level Diversity

Rather than focusing on the redundancy of recommended items to each customer, system-level diversity reflects the ability of the entire system to recommend those less popular or hard-to-find items. Under this category, all customers are aggregated all together at a system level and the diversity measures the dissimilarity among all the recommended items the entire system had made. A high degree of system-level diversity now indicates that the system can recommend a wide range of items rather than only those best-sellers or popular items, and is especially beneficial to those minority provider groups. In other words, system-level diversity is relevant to exposure fairness among providers, which is important for maintaining a healthy marketplace.

It is worth noting that individual-level diversity and system-level diversity address two distinct concerns with little overlap. System-level diversity is not a simple average of individual-level diversity across all customers. It is conceivable for a system to have a high degree of individual-level diversity but a low degree of system-level diversity, and vice versa. For example, if the system recommends to all

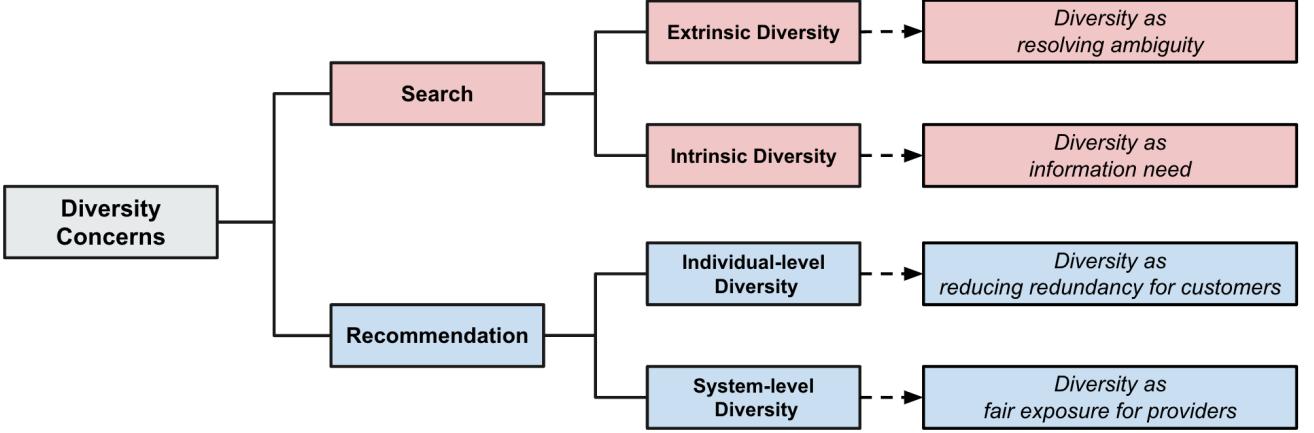
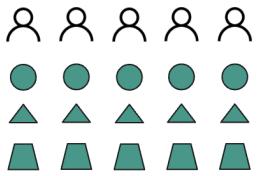
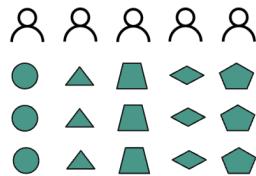


Fig. 2: Diversity in search and recommendation. Pink boxes indicate that they were originally proposed and generally used in search, while blue boxes indicate that they are generally used in the recommendation.



individual-level diversity: high
system-level diversity: low



individual-level diversity: low
system-level diversity: high

Fig. 3: A toy example to show that the individual-level diversity and system-level diversity in the recommendation are different concerns with little overlap. In this illustration, different shapes refer to different categories. Consider a top-3 recommendation and assume that there is an extremely large number of users and items in the system (the same as the real-world scenarios). In case 1, the system recommends the same 3 categories of items to each user; in case 2, the system always recommends a unique category of items to each unique user. Therefore, in case 1, the individual-level diversity is high and the system-level diversity is low, while in case 2, the individual-level diversity is low and the system-level diversity is high.

users the same five best-selling items that are not similar to each other, the recommendation list for each user is diverse (i.e., high individual-level diversity), but only five distinct items are recommended to all users and purchased by them (i.e., resulting in low system-level diversity or high sales concentration). In the other case, if the system recommends the same and unique category of items to each user, then the individual-level diversity is low, while the system-level diversity can be high. A toy example is provided in Fig. 3.

4 NOTATIONS

The notations we used in this paper are shown in Table 1.

5 METRICS OF DIVERSITY IN SEARCH AND RECOMMENDATION

Although many diversity metrics were proposed separately in either the field of search or recommendation, they can actually be applied interchangeably in both fields, since they all commonly aim to measure the dissimilarity and non-redundancy among a list of items. Coming up with a unified

TABLE 1: Description of Notations.

Notation	Description
\mathcal{U}, \mathcal{D}	The set of all users and items
\mathcal{D}_u	The set of interacted items of user u
u, d	An individual user / item
Θ	Learnable embeddings of users, items, and subtopics (if applicable)
\mathbf{M}	The interaction matrix between users and items
σ	A ranking list of items
$\sigma^{i:j}$	The list of items from position i to position j extracted from σ
$\sigma^{-1}(d)$	The ranking position of item d in σ
n_S	The total number of subtopics (categories)
$S(d)$	The set of subtopics covered by item d
c_s^l	The number of items covering subtopic s in list l
e_i	The exposure of item d_i in the entire system
$o(d u)$	The score of item d with respect to user u
$p(s u)$	The user u 's interest in subtopic s
$p(d s)$	The relatedness of items d to subtopic s

classification rationale for diversity metrics is necessary but not easy since diversity can be measured from multiple perspectives in search and recommendation. For example, some [19, 20] adopt the dissimilarity to define diversity and use the Intra-List Average Distance (ILAD) to measure diversity. Others [17, 21] consider the item position and relatedness to different subtopics to determine the diversity of the recommendation list and use α -NDCG to measure diversity. Some other researchers [22, 23] define diversity in a way where the relevance also being inherently considered. For instance, Agrawal et al. [22] state the problem of result diversification as: “Suppose users only consider the top k returned results of a search engine. Our objective is to maximize the probability that the average user finds at least one useful result within the top k results”. The word “useful” here is related to relevance inherently. The diversity metrics Agrawal et al. [22] later defined (i.e., the Intent-aware family (IA-family) of metrics) are also related to relevance themselves, since the metrics have captured user preference on recommended items based on the user intents.

In this survey, motivated by prior analysis and studies [14, 23, 24], we summarize the metrics of diversity in both fields under one unified taxonomy as follows. We first categorize diversity metrics into two classes based on whether the relevance of items¹ to the user (query) is taken into consideration: (i) **Relevance-oblivious Diversity Metrics** and (ii) **Relevance-aware Diversity Metrics**. We

1. To clarify, the term “item” in the rest of this paper can refer to both “entities retrieved from search systems” and “goods displayed by recommendation systems”.

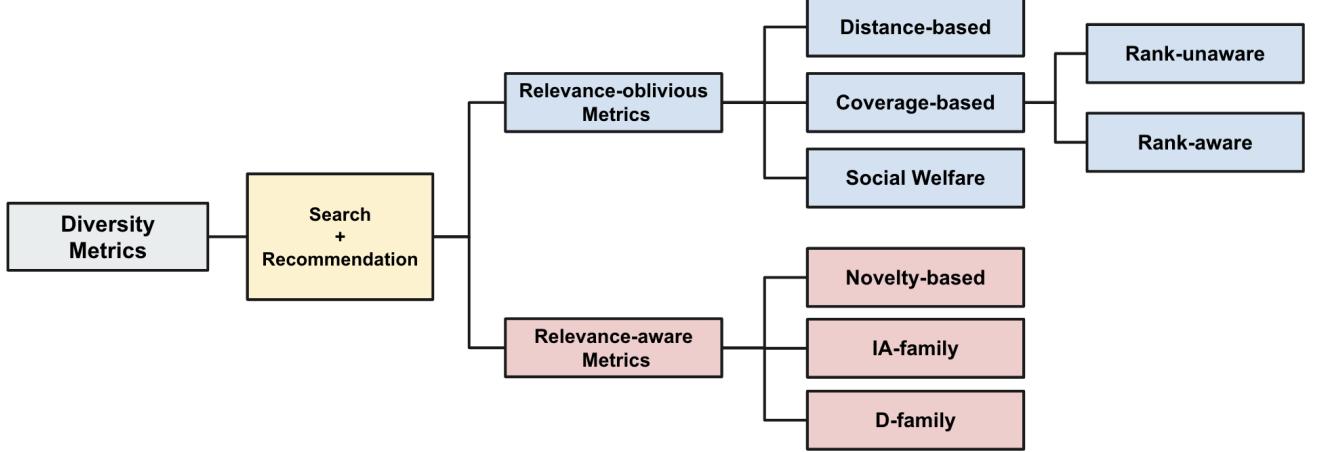


Fig. 4: Diversity metrics in search and recommendation. We unify the metrics in one unified classification since all metrics can be theoretically used in both fields. Metrics in pink boxes indicate that they were originally proposed and generally used in search, while those in blue boxes indicate that they are generally used in recommendation.

further classify these metrics into sub-classes in a more fine-grained manner. A summary table is maintained to highlight the applicability of all these metrics in either search, recommendation, or both. These metrics and corresponding works are summarized in Table 2.

5.1 Relevance-oblivious Diversity Metrics

The relevance-oblivious metrics do not take the relevance between items and user (query) into consideration, while merely focus on the diversity measurement of a ranking list itself. We further categorize these metrics into the following three sub-classes: *Distance-based Metrics*, *Coverage-based Metrics*, and *Social Welfare Metrics*.

5.1.1 Distance-based Metrics

One of the most widely used metrics for measuring diversity is the distance-based metric. As the name says, it evaluates the diversity by calculating the pair-wise distance among all the items in a list, where a smaller distance value indicates a poorer diversity. Given a specific criterion for computing the pair-wise distance, most works follow the ILAD and ILMD (short for Intra-List Average Distance and Intra-List Minimal Distance) paradigm for obtaining the diversity value of the item list. These two paradigms originated from paper [19] for measuring the diversity of a list, which extended the metrics of intra-list similarity proposed by Ziegler et al. [20]. We denote σ as a retrieval or recommendation list, dis_{ij} as the distance between item d_i and d_j . Then the ILAD can be defined as the average dissimilarity of all pairs of items in the list, while the ILMD can be defined analogously.

$$\text{ILAD} = \frac{1}{\binom{n}{2}} \sum_{i,j \in \sigma, i \neq j} \text{dis}_{ij}, \quad \text{ILMD} = \min_{i,j \in \sigma, i \neq j} \text{dis}_{ij}. \quad (1)$$

In some specific applications such as sequential recommendation, items are displayed as a sequence to the user, and the diversity is only required for w successive items [25]. We denote $\sigma^{-1}(d)$ as the rank position of item d in σ , then we can obtain two variants of ILAD and ILMD as ILALD

and ILMLD (short for Intra-List Average Local Distance and Intra-List Minimal Local Distance), which can be defined as:

$$\text{ILALD} = \frac{\text{mean}_{\substack{d_i, d_j \in \sigma, i \neq j, \\ |\sigma^{-1}(d_i) - \sigma^{-1}(d_j)| \leq w}} \text{dis}_{ij}}{w}, \quad (2)$$

$$\text{ILMLD} = \min_{\substack{d_i, d_j \in \sigma, i \neq j, \\ |\sigma^{-1}(d_i) - \sigma^{-1}(d_j)| \leq w}} \text{dis}_{ij}. \quad (3)$$

Based on different ways to calculate the pair-wise distance dis_{ij} , several specific metrics are summarized below.

- **Cosine Diversity Distance.** The most traditional and widely adopted way for defining the pair-wise distance is to use the cosine similarity between item embeddings, where dis_{ij} is generally defined as $\text{dis}_{ij} = 1 - \cos(\vec{d}_i, \vec{d}_j)$, where $\cos(\cdot, \cdot)$ refers to the cosine similarity. One of the primary advantages of cosine similarity is its simplicity, especially for sparse vectors — only non-zero entries need to be considered. This is also how the original ILAD and ILMD proposed by Ziegler et al. [20] define the pair-wise distance. After computing the cosine distance between any pair of items within the list, the Cosine diversity distance of the whole list can be obtained using the paradigm in Eq. 1.
- **Jaccard Diversity Distance.** Proposed by Yu et al. [75], the Jaccard diversity distance is calculated similarly to a standard Jaccard index paradigm. However, the exact distance is not computed based on item embeddings, but relies on *explanation* between user-item pairs. The *explanation* is defined differently given different recommendation models. If an item d_i is recommended to user u by a content-based strategy, then an *explanation* for recommendation d_i is defined as:

$$\text{Expl}(u, d_i) = \{d_j \in \mathcal{D} | \text{sim}(d_i, d_j) > 0 \wedge d_j \in \mathcal{D}_u\}. \quad (4)$$

Thereafter, the Jaccard diversity distance (JDD) between two items recommended to a specific user can be defined using the pre-computed *explanation*:

$$\text{JDD}(d_i, d_j | u) = 1 - \frac{|\text{Expl}(u, d_i) \cap \text{Expl}(u, d_j)|}{|\text{Expl}(u, d_i) \cup \text{Expl}(u, d_j)|}. \quad (5)$$

Then the Jaccard diversity distance of the whole list can be defined as Eq. 1.

TABLE 2: A lookup table for the publications proposing or using different diversity metrics in search and recommendation.

Diversity Metrics			Related Work	
Relevance-oblivious Diversity Metrics	Distance-based	Cosine Diversity Distance		[19, 20, 21, 25, 26, 27, 28, 29, 30, 31, 32] [33, 34, 35, 36, 37]
		Jaccard Diversity Distance		[38, 39]
		Gower Diversity Distance		[40]
	Coverage-based	Rank-unaware	P-Coverage	[35, 41, 42, 43, 44, 45, 46]
			C-Coverage	[37, 41, 42, 47, 48, 49]
			S-Coverage	[21, 31, 42, 50, 51]
			E-Coverage	[52, 53]
		Rank-aware	S-RR@100%	[54]
			S-Recall@K	[37, 54, 55, 56, 57]
			S-Precision@K	[54]
	Social Welfare	SD Index		[50, 58]
		Gini Index		[49, 59, 60, 61]
Relevance-aware Diversity Metrics	Novelty-based	α -nDCG@K		[17, 21, 34, 50, 55, 56, 62, 63, 64, 65, 66] [37, 57, 67]
		NRBP		[55, 57, 66, 67]
		nDCU@K		[68, 69]
	IA-family	M-IA		[22, 34, 55, 56, 57, 63, 64, 65, 66, 67, 70]
	D-family	D-M, D#-M		[56, 70, 71, 72, 73, 74]

- **Gower Diversity Distance.** Another metric belonging to the distance-based category is the Gower diversity distance, proposed by Haritsa [40], focusing on retrieving K nearest and diversified items with respect to a given query.

Motivated by the Gower coefficient [76], they define the distinction between two items as a weighted average of the respective attribute value differences. Assuming δ_k is the difference of the k^{th} attribute between two items and w_k is the corresponding weight on that attribute, then the Gower diversity distance (GDD) between two items d_i and d_j can be defined as:

$$\text{GDD}(d_i, d_j) = \sum_k w_k \cdot \delta_k(d_i, d_j). \quad (6)$$

The rest of the computation over a whole list follows the same paradigm in Eq. 1.

5.1.2 Coverage-based Metrics

Coverage-based metrics, popular for diversity measurement in search and recommendation, are often designed to quantify the breadth of *subtopics*² within a list of unique items. Depending on whether item ranks matter, we classify metrics as *rank-unaware* or *rank-based*. The first category disregards the ranking positions of items.

Rank-unaware. Rank-unaware metrics are similar to the conventional metrics on accuracy in search and recommendation (e.g., Precision@K and Recall@K) since both of them will not be influenced by the rank positions of items in a given list. Depending on the coverage of “what” they measure, we can classify these metrics into three sub-classes: P-Coverage, C-Coverage, and S-Coverage.

- **P-Coverage** (short for Prediction Coverage). The measure of prediction coverage is the number of unique items for which the predictions can be formulated as

2. The term “*subtopic*” originates from information retrieval, indicating the presence of multiple themes or keywords relevant to the input query. For consistency, we use “*subtopic*” to represent (i) *category of items*, (ii) *aspect of queries*, and (iii) *intent of users* in this survey.

a proportion of the total number of items [41, 42]. We denote \mathcal{D} as the set of all available items, \mathcal{D}_p as the set of items for which a prediction can be provided. Then the P-Coverage can be defined as follows:

$$\text{P-Coverage} = \frac{|\mathcal{D}_p|}{|\mathcal{D}|}. \quad (7)$$

The construction of \mathcal{D}_p is highly dependent on the task formulation and chosen models. For instance, paper [42] mentioned that some collaborative filtering systems are just able to make predictions for items that have more than a fixed number of ratings assigned. In such a case, \mathcal{D}_p can be considered as the set of items for which the number of available ratings meets the requirement. P-Coverage generally focuses on the system level, and is hardly used for measurement on a single or several ranking lists.

From another view, we can also understand the P-Coverage as the system’s ability to address the “cold-start” problem. However, as more and more research on “cold-start” problem emerges, most models are capable of making predictions for those items even with very few interactions. Thus, P-Coverage is not widely used in search and recommendation.

- **C-Coverage** (short for Catalog Coverage). In order to quantify the proportion of unique items that can be retrieved or recommended in the system, the catalog coverage metric directly focuses on the output result list and generally considers the union of all lists produced during the measurement time [41, 42]. C-Coverage can be used to measure the diversity of either a single ranking list or a group of lists, but is more widely used at a system level. Assuming there are N lists, the metric can be formulated as follows, where $\text{set}(\cdot)$ is to convert a ranking list to a set:

$$\text{C-Coverage} = \frac{\left| \bigcup_{i=1}^N \text{set}(\sigma_i) \right|}{|\mathcal{D}|}. \quad (8)$$

- **S-Coverage** (short for Subtopic-Coverage). This metric is one of the most widely used measurements for diver-

sity in search and recommendation [42, 51]. Differing from C-Coverage, which focuses on the items themselves, S-Coverage considers the variety and richness of different item categories or genres within the list. This aligns more naturally with human perceptions than distance-based metrics, as people typically do not compute pair-wise distances based on embeddings to assess list diversity, but rather identify whether diverse topics are as frequent as possible. S-Coverage can be gauged on either a single list or multiple lists, corresponding respectively to individual-level and system-level diversity measurements. If we denote N as the number of lists in consideration, $\mathcal{S}(d)$ as the set of subtopics covered by item d , and $n_S = |\bigcup_{d \in \mathcal{D}} \mathcal{S}(d)|$ as the total number of subtopics, then S-Coverage can be expressed as follows:

$$\text{S-Coverage} = \frac{\left| \bigcup_{i=1}^N (\bigcup_{d \in \sigma_i} \mathcal{S}(d)) \right|}{n_S}. \quad (9)$$

- **Density** is another member in the family of coverage-based metrics, which derives from the network science and is widely applied on graphs and information networks [52, 53]. The density of a graph is defined as the number of edges (excluding self-links) presenting in the network divided by the maximal possible number of edges in the network. We re-name it as **E-Coverage** (short for Edge-Coverage) in our survey.

Rank-aware. It has been realized in many works that users do not provide all items in a ranking list with the same amount of attention due to users' patience may decay exponentially as they browse deeper through a list. As a result, those items ranked higher (i.e., at the top of the list) may receive more exposure. Thus, when considering relevance, many metrics have been proposed to offer a higher score for ranking relevant items at the top of a list, such as the normalized discounted cumulative gain (nDCG).

A similar idea is also applicable when considering the diversity of a list: a user may feel the list is redundant if those items ranked in top positions are similar to each other, even if there are many diverse items in later positions. This nuance cannot be captured by prior described metrics since they are invariant when the ranks of items change. Here, we summarize the rank-based metrics on measuring coverage-based diversity, where most of them are defined upon the conventional metrics for measuring accuracy in search and recommendation. These metrics care about not only how diverse the items are but also what locations they appear.

- **S-RR@100%** (short for Subtopic-Reciprocal Rank@100%). This metric is proposed in [54] for evaluating the diversity of solutions for subtopic retrieval problems. The subtopic retrieval problem is originally concerned with finding documents that cover many different subtopics given a query of keywords. S-RR@100% is a variation to Reciprocal Rank (RR), defined as the inverse of the rank position on which a complete coverage of subtopics is obtained. Thus, the output value of this metric cannot be smaller

than the total number of different subtopics. Using the same notation as before, S-RR@100% can be defined as:

$$\text{S-RR@100\%} = \min_k \left(\left| \bigcup_{i=1}^k \mathcal{S}(d_i) \right| = n_S \right). \quad (10)$$

- **S-Recall@K** (short for Subtopic-Recall@K). As the name says, this metric is a variation of the Recall@K metric that is widely used for measuring relevance in search and recommendation. S-Recall@K is also proposed in [54] and can be defined as the percentage of subtopics covered by the first k items given a retrieved or recommendation list:

$$\text{S-Recall@K} = \frac{\left| \bigcup_{i=1}^k \mathcal{S}(d_i) \right|}{n_S}. \quad (11)$$

- **S-Precision@K** (short for Subtopic-Precision@K). Analogous to S-Recall@K, this metric is a variation of the Precision@K. It can be defined as below:

$$\text{S-Precision@K} = \frac{\left| \bigcup_{i=1}^k \mathcal{S}(d_i) \right|}{k}. \quad (12)$$

5.1.3 Social Welfare Metrics

Diversity is not only a research problem of information retrieval in computer science. Additionally, it has received lots of attention in other disciplines such as ecology and economics. Recently, several works borrow the diversity notions from other fields for evaluating search and recommendation results. We summarize these metrics as follows.

- **SD Index** (short for Simpson's Diversity Index). SD Index originated from paper [58] for measuring the biodiversity in a habitat. Regarding each subtopic (category) in recommendation as a kind of species in ecology, SD Index can be defined as the probability that two items selected randomly and independently without replacement belong to the same category. Thus, a smaller SD Index indicates a higher diversity. We denote n_S as the number of different subtopics, l as the list of items under consideration (which can be a single recommendation list or the concatenation of multiple lists), and $c_{s_i}^l$ as the number of items covering the subtopic s_i in the list l . Then the SD Index over the list l can be defined as follows:

$$\text{SD Index} = \frac{\sum_{i=1}^{n_S} [c_{s_i}^l \cdot (c_{s_i}^l - 1)]}{|l|(|l| - 1)}. \quad (13)$$

To better demonstrate this, we assume 3 subtopics in total. System *A* recommends 10 items and the number of items covering each subtopic is 8, 1, 1. System *B* also recommends 10 items, while the number of items covering each subtopic is 4, 3, 3. Then we can compute the SD Index of both systems and find out that the index value of system *A* is larger than that of system *B*: $\frac{8 \times 7 + 1 \times 0 + 1 \times 0}{10 \times 9} > \frac{4 \times 3 + 3 \times 2 + 3 \times 2}{10 \times 9}$. This means that system *A*'s recommendation is less diverse than system *B*'s recommendation, which is aligned with our intuition.

- **Gini Index**. The Gini Index proposed by Gini [61] is originally a measure of the distribution of income across a population. A higher Gini Index indicates

greater inequality, with high-income individuals receiving much larger percentages of the total income of the population. Recently, some researchers also adopt the Gini Index in the field of recommendation to measure the inequality among values of a frequency distribution, e.g., the number of occurrences (exposures) in the recommendation list. This measurement is generally at the system level by aggregating all the recommendation lists across all users, which can also indicate how diverse the system is in regard to all the items it can retrieve or recommend. Assuming the occurrence of the i^{th} item is e_i , where $i = 1, \dots, |\mathcal{D}|$, the Gini Index over all the items of the whole system is calculated as:

$$\text{Gini Index} = \frac{1}{2|\mathcal{D}|^2\bar{e}} \sum_{i=1}^{|\mathcal{D}|} \sum_{j=1}^{|\mathcal{D}|} |e_i - e_j|, \quad (14)$$

where $\bar{e} = \frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} e_i$ is the mean occurrence of items. Thus, a smaller Gini Index indicates a more fair distribution of the occurrences of items in the output results. This may indicate a higher diversity since different items have more equal opportunities to be exposed.

5.2 Relevance-aware Diversity Metrics

Although diversity is an important property algorithm designers need to consider, relevance is still at the heart of the ranking problems in both search and recommendation. Therefore, a metric that is solely concerned with diversity cannot adequately assess a system's effectiveness. For instance, one can randomly select items from various topics to ensure that the ranking list performs exceptionally well on metrics such as S-Coverage and S-RR@100%. However, the overall relevance of the list obtained in such a way may be extremely low. Therefore, if the algorithm designer aims to use the relevance-oblivious metrics to measure the diversity, she has to use other metrics (e.g., nDCG, rank-biased precision (RBP) [77]) to measure the relevance.

In contrast to those relevance-obvious metrics on diversity, there exist relevance-aware metrics that attempt to incorporate both relevance and diversity into a single measurement, where almost all of them originated from research in search. Several works conduct axiomatic analysis on the relevance constraints of diversity metrics [78, 79]. Here, we highlight two of the most critical properties on the relevance in ranking, **priority** and **heaviness**, that the relevance-aware metrics in any ranking task must satisfy.

- **Property 1: Priority.** Swapping items in concordance with their relevance scores should increase the overall score of the whole ranking.

We denote $o(d)$ as the relevance score of an item with respect to the query or user, $Q(\sigma)$ as the overall score of the ranking list σ , “ \leftrightarrow ” as swapping two items. Formally, the **priority** property requires that: if $o(d_i) < o(d_j)$ and $\sigma^{-1}(d_i) < \sigma^{-1}(d_j)$, then $Q(\sigma_{d_i \leftrightarrow d_j}) > Q(\sigma)$.

- **Property 2: Heaviness.** Items with the same relevance score contribute more when at earlier positions in the ranking.

Using the same notations, the **heaviness** property requires that: if $o(d_i) = o(d_j) < o(d_{i'}) = o(d_{j'})$, $\sigma^{-1}(d_{i'}) - \sigma^{-1}(d_i) > 0$, $\sigma^{-1}(d_j) - \sigma^{-1}(d_i) = \sigma^{-1}(d_{j'}) - \sigma^{-1}(d_{i'}) > 0$, then $Q(\sigma_{d_i \leftrightarrow d_{i'}}) > Q(\sigma_{d_j \leftrightarrow d_{j'}})$.

It is easy to see that those widely used relevance metrics in information retrieval and recommendation, such as the

nDCG, satisfy both of the two properties. Now, we categorize the relevance-aware diversity metrics that satisfy the above properties into the following two categories.

5.2.1 Novelty-based Metrics

As a metric with a close connection to diversity, novelty was also studied in prior works. Clarke et al. [17] point out the precise distinction between novelty and diversity in the field of information retrieval: *novelty refers to the need to avoid redundancy, while diversity refers to the need to resolve ambiguity in the query*, which corresponds to *intrinsic diversity* and *extrinsic diversity*, respectively. However, even with this difference, we still find out several works in the literature categorize the novelty-based metrics as part of the metrics in the diversity family, since novelty on topics and categories can also be regarded as an improvement on diversity. We follow this paradigm and summarize the novelty-based metrics as follows.

- α -nDCG@K (short for Novelty-biased Normalized Discounted Cumulative Gain@K). This metric is proposed by Clarke et al. [17], using to balance the trade-off between retrieving both relevant and non-redundant items. This is also one of the earliest metrics aiming to combine the measurement of both relevance and diversity. Prior to this, most metrics in information retrieval and recommendation such as mean average precision (MAP) and nDCG assume that the relevance of each item can be judged in isolation, independently from other items, thus ignoring important factors such as redundancy between items. To address this, the authors present a framework for assessing diversity and novelty based on the cumulative gain.

The key is to define the utility gain of adding the k^{th} item (d^k) in the list to be considered in the left of all items ranked above position k . The authors assume that subtopics are independent and equally likely to be relevant, and the assessment of positive relevance judgments of an item for a subtopic $o(d|s)$ involves an uncertainty that can be modeled with a fixed probability α of success in the judgment.

We denote $c_{s_i}^{\sigma^{1:k}}$ as the number of items covering subtopic s_i till position k in the ranking list σ , then we can first formally define the α -DCG@K over a ranking list σ as below, where $0 < \alpha \leq 1$ [17]:

$$\text{Gain}(d^k) = \sum_{i=1}^{n_S} o(d^k|s_i)(1 - \alpha)^{c_{s_i}^{\sigma^{1:k}}}, \quad (15)$$

$$\alpha\text{-DCG@K} = \sum_{k=1}^K \frac{1}{\log(k+1)} \cdot \text{Gain}(d^k). \quad (16)$$

Analogous to the definition of nDCG@K, we can find an “ideal” ranking that maximizes α -DCG@K, denoting as the α -iDCG@K. The ideal ranking computation is known to be an NP-Complete problem, pointed out by Carterette [80]. The ratio of α -DCG@K to α -iDCG@K defines the α -nDCG@K.

- **NRBP** (short for Novelty- and Rank-Biased Precision). Following the paper [17], Clarke et al. [81] propose another metric built upon the rank-biased precision (RBP), rather than the nDCG, with a very similar motivation

and paradigm. Described by Moffat and Zobel [77], the RBP model assumes that the user browses a list of items in sequential order and with a probability β (i.e., $0 < \beta < 1$) to continue at each position. In other words, a user has a probability of β^k to observe all the items till the k^{th} position. Based on this idea and using the same notation as α -nDCG@K in Eq. 16, the NRBp can be formally defined as follows:

$$\text{NRBP} = \frac{1 - (1 - \alpha)\beta}{n_S} \sum_{k=1}^{|\sigma|} \beta^{k-1} \sum_{i=1}^{n_S} o(d^k | s_i) (1 - \alpha)^{c_{s_i}^{\sigma^{1:k}}}. \quad (17)$$

Here, the normalization factor includes division by the number of subtopics, allowing us to average the measure across multiple queries with varying subtopic counts. It is also worth mentioning that in contrast to α -nDCG@K which is typically presented at a particular browsing depth, NRBP is more of a summary metric that illustrates the diversity/novelty of the entire list.

- **nDCU@K** (short for Normalized Discounted Cumulative Utility@K). The nDCU@K was proposed by Yang et al. [68] around the same period when the α -nDCG@K was proposed. It is also motivated by extending the original nDCG@K metric. Given a list of retrieved items based on a query, the authors define the utility of an item d^k at the k^{th} position as:

$$\text{Utility}(d^k) = \text{Gain}(d^k) - \text{Loss}(d^k), \quad (18)$$

where the $\text{Gain}(d^k)$ refers to the information users receive for observing d^k depending on the relevance and novelty, while $\text{Loss}(d^k)$ denotes the time and energy spent in going through the item. There are various ways to define these two terms. Yang et al. [68] adopt a very similar formulation as that in Eq. 15 to define the $\text{Gain}(d^k)$, and they define the $\text{Loss}(d^k)$ as a constant. Then the DCU@K can be defined as:

$$\text{DCU@K} = \sum_{k=1}^K \frac{1}{\log(k+1)} \cdot \text{Utility}(d^k). \quad (19)$$

Analogous to the nDCG@K, the ratio of DCU@K to the ideal DCU@K defines the nDCU@K.

Considering that nDCU@K is well-defined only for a single ranking list [68], Yang and Lad [69] extend the nDCU@K metric to make it be capable of measuring multiple ranking lists conditioned on different user browsing patterns. Specifically, they compute the mathematical expectation of Eq. 18 over n ranking lists and n user browsing models (one list and one browsing model for each user), where each browsing model corresponds to a list of n different stop positions in the n ranking lists. Since this is just a slight modification of nDCU@K, we do not treat it as a separate metric in this survey.

5.2.2 IA-family Metrics

Sometimes, a simple combination of diversity and per-intent graded relevance is still not enough. When considering diversity, it is not always ideal to retrieve or recommend different items covering various topics without distinguishing which topic is more important. Take the search problem as an example, “intent” is defined as the “information

needs”, which can also be referred as users’ expectations on distributions of different “subtopics” in the search result with respect to a query.

The motivation behind intent-aware metrics can also be depicted by the following example described in paper [22]. Considering a query q that is related to two subtopics s_1 and s_2 , but is much more related to s_2 than s_1 . Now we have two items d_1 and d_2 , where d_1 rated 5 (out of 5) for s_1 but unrelated to s_2 , d_2 rated 4 (out of 5) for s_2 but unrelated to s_1 . Traditional relevance metrics tend to rank d_1 over d_2 , but users may find d_2 is more related than d_1 on average.

As such, Agrawal et al. [22] propose the family of intent-aware metrics for search results diversification. Formally, given a distribution on the subtopics for a query and a relevance label on each item, they compute the outcome over a list σ by applying the intent-aware scheme on a conventional relevance metric M as follows:

$$M\text{-IA}(\sigma) = \sum_{i=1}^{n_S} p(s_i | q) \cdot M(\sigma | s_i), \quad (20)$$

where s is the subtopic, denoting the user intent. M is the traditional metric for measuring the ranking quality on relevance, such as nDCG, MRR, and MAP. When computing the intent-dependent $M@K(\sigma | s)$, they simply treat any item that does not match the subtopic s as non-relevant items, then compute the same way as the original $M@K(\sigma)$. Thus, the family of $M\text{-IA}$ metrics takes into account the distributions of intents, and force a trade-off between adding items with higher relevance scores and those that cover intents with higher weights.

5.2.3 D-family Metrics

The intent-aware metrics are sub-optimal in that they are not guaranteed to range between 0 and 1: it is generally not possible for a single system output to be ideal for all intents at the same time. As such, Sakai and Song [72], Sakai et al. [73] propose an alternative way to evaluate diversified search results, given intent probabilities and per-intent graded relevance assessments. While intent-aware measures combine multiple measure scores using the intent probabilities, the D-family combines per-intent relevance grades of each document using the intent probabilities.

Given the intent probabilities $p(s_i | q)$ and per-intent graded relevance assessments, where $g(d^k | s_i)$ is the gain value for document at rank position k for intent s_i , the global gain at rank position k can be defined as:

$$\text{GG}(k) = \sum_{i=1}^{n_S} p(s_i | q) \cdot g(d^k | s_i). \quad (21)$$

For a launched query, an ideal ranking list can be obtained by sorting all documents by the global gain. Denoting the global gain of the document ranked at position k in the retrieval list and the ideal ranking list, D-metrics can be computed by using these gain values instead of the traditional pre-set values.

Apart from the fact that D-metrics avoid the undernormalisation problem of intent-aware metrics by relying on a single “globally ideal” list, these two metric families are quite similar. However, Sakai and Song [72], Sakai et al. [73] also propose a simple method to explicitly encourage high

intent recall (S-Recall) in a search output within the D-metric framework. Then D#-metrics are defined as:

$$\text{D\#-M@K} = \gamma \cdot \text{S-Recall@K} + (1 - \gamma) \cdot \text{D-M@K}. \quad (22)$$

where γ is a parameter, M is the traditional metric for measuring the ranking quality on relevance, such as nDCG, MRR, and MAP. Several other works such as [74] propose further extensions of D/D#-metrics by introducing intent hierarchies to model the relationships between intents, and present various weighing schemes. In this survey, we do not classify them separately and still treat them as instances of the D-family metrics.

6 OFFLINE APPROACHES FOR ENHANCING DIVERSITY

We intend to use a unified framework to categorize the approaches for enhancing diversity in both search and recommendation since there are lots of similarities in these two scenarios. In this section, we focus on offline processes, where the methods do not need to care about users' real-time feedback based on the displayed results from the last round. Based on when the approaches of diversity intervene relative to the training procedure, the offline diversity approaches can be divided into three categories: (i) **pre-processing**, (ii) **in-processing**, and (iii) **post-processing**. Pre-processing methods are adopted prior to the model training process. They typically pre-defined diversity-aware techniques with the expectation that these designs will result in a diverse output. In-processing methods directly participate during the model training. They guarantee a diversified outcome through learning matching scores for users and items where the diversity constraints or scores are added. Post-processing methods are used after the model is well-trained. They generally re-rank the final item lists to improve diversity. All approaches and corresponding works are summarized in Table 3.

6.1 Pre-processing Methods

Pre-processing methods intervene in the system before the model training. We review the following three sub-classes (i) *pre-defined user types*, (ii) *pre-defined sampling strategies*, and (iii) *pre-defined ground-truth label*.

6.1.1 Pre-defined User Types

Kwon et al. [48] improve the diversity in recommendation through pre-defining user types. They first interview fashion professionals and categorize the user purchase behavior into four types: (i) *gift type*: purchasing items to give to others; (ii) *coordinator type*: purchasing items associated with previous purchases; (iii) *carry-over type*: purchasing items similar to existing purchases; and (iv) *trend-setter type*: purchasing items affected by the trends of other people's purchases. For each type, they use a specific algorithm to recommend top-5 items to each user. Since each user may have multiple purchase behaviors, they adopt a hybrid strategy to simply combine the recommendation lists of four algorithms to generate up to 20 item candidates for each user. The C-Coverage improves from 24% (using the CF algorithm) to 78%, coupled with a 3.2% increase in the purchasing rate and a \$13 increase in the average purchase amount per customer in the experimental group.

6.1.2 Pre-defined Sampling Strategies

Since the interactions between users and items can be naturally represented as a graph, a number of graph-based recommendation algorithms are widely used nowadays, such as the Graph Neural Network (GNN)-based models. These models represent the users and items as nodes in a graph, followed by learning their embeddings through message passing: at each layer, they aggregate the neighbor's information for each target node. Due to their ability to capture higher-order connectivity between user nodes and item nodes, GNN-based methods can generally achieve state-of-the-art accuracy and relevance.

Since higher-order neighbors of a user tend to cover more diverse items, GNN-based approaches have the potential to improve recommendation diversity as a byproduct. Without specific design, those items from the popular categories tend to be learned more often, because they take up the majority of the edges on the graph. To address this, Zheng et al. [49] propose two pre-defined sampling strategies for two processes in the model. The first strategy aims to re-balance the *neighborhood sampling* process in the message passing to increase the selecting probabilities for those items from the less popular categories and decrease those from the popular categories. In such a way, those less popular items can still have a chance to be sampled and well-learned. The second strategy affects the *negative sampling* process. In contrast to random negative sampling in paper [1], they propose to select similar (from the same category) but negative items with an increased probability, so that less similar (not from the same category) items are not pushed too far away from the user in the embedding space. As a result, items from different categories are likely to appear in the recommendation list for each user, thus enhancing the individual-level diversity. An illustration of this category-boosted negative sampling is shown in Fig. 6.

6.1.3 Pre-defined Ground-truth Label

Cheng et al. [21] construct ground truth labels via diversity constraints to directly idealize the optimization target. Employing supervised learning, each user becomes a training instance with a heuristically chosen subset of relevant, diverse items as their ground-truth label. Their two-step labelling method involves: (i) filtering high-rated items into a candidate set \mathcal{C}_u , and (ii) selecting items from \mathcal{C}_u to maximize the relevance-diversity balance. An item d qualifies as high-rated for user u if $o(d|u) \geq \gamma \cdot \bar{o}(\cdot|u)$, where $o(d|u)$ denotes the observed score of d by u , $\bar{o}(\cdot|u)$ is the average score across all items by u , and γ is a trade-off parameter.

All selected items in step (i) form the set of candidates \mathcal{C}_u . In step (ii), they select a subset \mathcal{Y}_u from \mathcal{C}_u (i.e., $\mathcal{Y}_u \subseteq \mathcal{C}_u$, $|\mathcal{Y}_u| = K$) as the ground-truth label for user u by balancing the trade-off between relevance and diversity, using a metric similar to the F-measure [95]. Specifically, the selected \mathcal{Y}_u aims to maximize the following equation:

$$\max_{\mathcal{Y}_u} \frac{2 \cdot f(\mathcal{Y}_u) \cdot g(\mathcal{Y}_u)}{f(\mathcal{Y}_u) + g(\mathcal{Y}_u)}, \quad (23)$$

$$\text{s.t., } \mathcal{Y}_u \subseteq \mathcal{C}_u, |\mathcal{Y}_u| = K. \quad (24)$$

Here, $f(\mathcal{Y}_u)$ and $g(\mathcal{Y}_u)$ represent the measurement for relevance and diversity over the whole set \mathcal{Y}_u , respectively. For

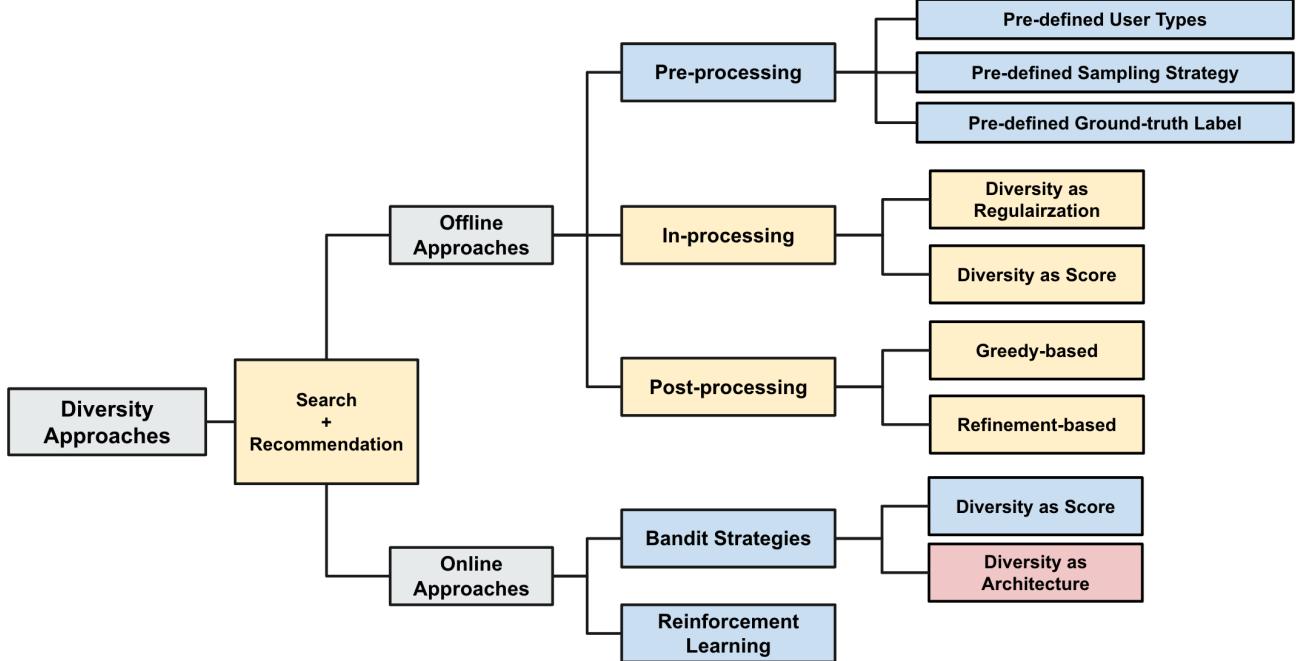


Fig. 5: Diversity approaches in search and recommendation, from both offline and online perspectives. Approaches in blue boxes indicate that they are generally used in recommendation, the pink box indicates it is generally used in search, while those in yellow boxes are equally widely used in search and recommendation.

TABLE 3: Summary for the publications proposing or using different diversity approaches in search and recommendation.

Diversity Approaches		Related Work	
Offline Approaches	Pre-processing Methods	Pre-defined User Types	[48]
		Pre-defined Sampling Strategies	[49]
		Pre-defined Ground-truth Label	[21, 43]
	In-processing Methods	Diversity as Regularization	[27, 33, 36, 44, 50, 60, 82, 83]
		Diversity as Score	[37, 55, 63, 65]
	Post-processing Methods	Greedy-based	[16, 34, 46, 53, 64, 84]
		MMR DPP	[25, 26, 29, 30, 38, 47, 85]
Online Approaches	Bandit Strategies	Refinement-based	[20, 28, 39, 45, 46, 52, 53, 66, 67, 86, 87]
		Diversity as Score	[88, 89, 90, 91]
	Reinforcement Learning	Diversity as Architecture	[92, 93]
			[35, 94]



Fig. 6: An illustration of the category-boosted negative sampling. Negative items are sampled from outside positive items. The strategy boosts the probability of sampling from items of positive categories (the light green area). The figure is borrowed from paper [49].

the relevance, denoting the set of items rated by u as \mathcal{D}_u , Cheng et al. [21] define $f(\cdot)$ as a pair-wise comparison on the ratings between the items in \mathcal{Y}_u and $\mathcal{D}_u \setminus \mathcal{Y}_u$ as follows:

$$f(\mathcal{Y}_u) = \frac{\sum_{d_i \in \mathcal{Y}_u} \sum_{d_j \in \mathcal{D}_u \setminus \mathcal{Y}_u} \text{compare}\left(o(d_i|u) - o(d_j|u)\right)}{|\mathcal{Y}_u| \cdot |\mathcal{D}_u \setminus \mathcal{Y}_u|}, \quad (25)$$

where $\text{compare}(x)$ equals 1 if $x > 0$; else, equals -1. For the diversity measurement $g(\cdot)$, the authors define it as the ILAD as described in Eq. 1. Afterward, the obtained ground-truth label \mathcal{Y}_u for user u can guide the model training.

6.2 In-processing Methods

In-processing methods act during the model training process. We categorize them into the following two sub-classes: (i) *diversity as regularization* and (ii) *diversity as score*.

6.2.1 Diversity as Regularization

Since relevance is the primary goal of search and recommendation systems, the most intuitive way to enhance diversity through an in-processing way is to treat diversity as a regularization on the loss function to guide the training. Wasilewski and Hurley [33] first propose a prototype to constrain the relevance loss with a trade-off parameter λ :

$$\min_{\Theta} \mathcal{L}_{\text{rel}}(\Theta) + \lambda \cdot \mathcal{L}_{\text{div}}(\Theta), \quad (26)$$

where Θ is the learnable embeddings, $\mathcal{L}_{\text{rel}}(\cdot)$ and $\mathcal{L}_{\text{div}}(\cdot)$ refer to the relevance loss and diversity regularization which can be both self-defined. For instance, Wasilewski and Hurley [33] define the relevance loss as the pair-wise ranking loss [1], and the diversity loss as the negative of the intra-list average distance (ILAD) [19].

Several later works follow this line of research, such as paper [27]. Rather than only modeling the dissimilarities

among items for defining the diversity, the authors take the user intents into consideration. Here, the user intents can be comprehended as the user's interest in different subtopics (categories). Specifically, the authors define the diversity of a recommendation list σ_u to user u as the probability of each subtopic s_i having at least one relevant item in σ_u , then the regularization can be formulated as:

$$\mathcal{L}_{\text{div}}^{\sigma_u} = - \sum_{i=1}^{n_S} p(s_i|u) \cdot \left(1 - \prod_{d \in \sigma_u} [1 - p(d|s_i)] \right). \quad (27)$$

Here, $p(s_i|u)$ represents u 's interest in subtopic s_i , and $p(d|s_i)$ refers to the relatedness of d to the subtopic s_i . Both terms can be computed through a softmax function using the embeddings of users, items, and subtopics.

In addition to merely focusing on the diversification of retrieval results, some researchers also care about how to generate diverse explanations for the output results. For instance, Balloccu et al. [44] conceptualize, assess, and operationalize three novel properties (linking interaction recency, shared entity popularity, and explanation type diversity) to monitor explanation quality at the user level in recommendation, and propose re-ranking approaches able to optimize for these properties. They optimize these three indicators for measuring the quality of explanations as a regularization term in the re-rank stage. Here we classify this as an in-processing method.

6.2.2 Diversity as Score

Another widely adopted in-processing method for diversity is to treat diversity as a score during ranking. As such, the score of an item is composed of two parts: one from the perspective of relevance, and the other from the perspective of diversity. The most significant difference between these two types of scores is that the relevance score typically assumes the independence of items in a list, but the diversity score is highly dependent on the other items.

Following this line, Li et al. [63] propose one of the earliest methods of diversified recommendation. They focus on a sequential recommendation process, where the model recommends one item at a time to form the entire recommendation list. They define the score of an item at the current position as the sum of a relevance part and a discounted subtopic diversification part. In detail, for user u , given an un-selected k^{th} item d^k and a list of selected $k-1$ items $\sigma_u^{1:k-1}$ (i.e., the list of first $k-1$ items in the ranking list σ_u), they define the score of d^k as:

$$o(d^k|u) = o^{\text{rel}}(d^k|u) + \lambda \cdot o^{\text{div}}(d^k|\sigma_u^{1:k-1}, u), \quad (28)$$

where $o^{\text{rel}}(d^k)$ and $o^{\text{div}}(d^k)$ denote the score of d^k from the view of relevance and diversity, respectively. Specifically, the authors define the relevance score as the inner product between the user and item embedding: $o^{\text{rel}}(d^k|u) = \Theta_u \cdot \Theta_{d^k}^\top$. They define the diversity score as discounted subtopic's contribution, which reduces exponentially as the number of items covering that subtopic increases in the entire list:

$$o^{\text{div}}(d^k|\sigma_u^{1:k-1}, u) = \sum_{i=1}^{n_S} \beta^{c_{s_i}^{\sigma_u^{1:k}}} \cdot \Theta_u \cdot \Theta_{s_i}^\top, \quad (29)$$

$$\sigma_u^{1:k} = \sigma_u^{1:k-1} \cdot \text{append}(d^k). \quad (30)$$

Here β is the decay factor (i.e., $0 < \beta < 1$), Θ_u is the embedding of user u , Θ_{s_i} is the embedding of subtopic s_i , $c_{s_i}^{\sigma_u^{1:k}}$ denotes the number of items covering subtopic s_i in $\sigma_u^{1:k}$. As such, for each user, the model greedily selects an un-selected item to maximize the score in Eq. 28 at each position to form the final recommendation list.

To train embedding Θ , Li et al. [63] assume that the ideal recommendation lists for some sample users are available. Then, the learning process aims to penalize those generated recommendations which do not respect the sequence in ideal lists. Taking an individual user u as an example, the loss function on a pair of sampled items (d_i, d_j) can be formulated through a binary cross-entropy loss:

$$\mathcal{L}_u(d_i, d_j) = -y_{ij} \cdot \log(p_{ij}) - (1 - y_{ij}) \cdot \log(1 - p_{ij}), \quad (31)$$

where $y_{ij} = 1$ if d_i is ranked above d_j in the ideal ranking list of u , p_{ij} refers to the probability of ranking d_i above d_j in the current model, which is computed as $p_{ij} = \text{sigmoid}(o(d_i|u) - o(d_j|u))$.

Other works also follow a similar idea to treat diversity as a score. For instance, Yu [65] presents a novel framework for search result diversification based on the score-and-sort method using direct metric optimization. They express each item's diversity score specifically, which determines its rank position based on a probability distribution.

6.3 Post-processing Methods

Earliest diversity approaches follow the re-ranking paradigm: they achieve diversity after the training procedure by re-ranking the list based on both relevance scores and diversity metrics. Due to the separation of model training and diversified ranking, these approaches are regarded as post-processing, which can be applied to any recommendation models as a consecutive layer with excellent scalability. Based on how the diversified list is generated, we categorize them as (i) *greedy-based* and (ii) *refinement-based*.

6.3.1 Greedy-based Methods

As the name suggests, greedy selection methods iteratively select the item that maximizes a joint measure of relevance and diversity to each position, and finally provide an output ranking list. Two of the most representative post-processing methods in this category are (i) *Maximal Marginal Relevance* (MMR) and (ii) *Determinantal Point Process* (DPP).

MMR Maximal Marginal Relevance (MMR) [16] is the most pioneering diversity approach in this category. Carbonell and Goldstein [16] propose "marginal relevance" as a linear combination of the relevance and diversity of each item, in response to the fact that user needs include not only *relevance* but also *novelty* and *diversity*. In particular, an item has high marginal relevance if it is both relevant to the user and has low similarity to previously selected items.

Based on this protocol, MMR greedily selects the item that can maximize the marginal relevance to form the final ranking list. We can formulate the process of MMR selecting the k^{th} item for user u as follows:

$$d^k = \max_{d \in (\mathcal{D} \setminus \mathcal{D}_u) \setminus \text{set}(\sigma_u^{1:k-1})} [o^{\text{rel}}(d|u) + \lambda \cdot o^{\text{div}}(d|\sigma_u^{1:k-1}, u)], \quad (32)$$

where $\text{set}(\sigma_u^{1:k-1})$ is the set of items composed by the first $k-1$ items in the ranking list σ_u .

Different researchers adopt similar ways to model $o^{\text{rel}}(\cdot)$ as the inner product between the embeddings, while adopting different ways to model the $o^{\text{div}}(\cdot)$. Carbonell and Goldstein [16] define the diversity term as:

$$o^{\text{div}}(d|\text{set}(\sigma_u^{1:k-1}), u) = - \max_{d_j \in \text{set}(\sigma_u^{1:k-1})} \text{sim}(d, d_j). \quad (33)$$

One may observe that the recommendation generation process of MMR is quite similar to that of paper [63], which falls under the category “In-processing - Diversity as Score”. The difference between these two works is as follows. MMR is a post-processing method that performs a greedy selection based on already learned model embeddings. In contrast, paper [63] adopts an in-processing method, whose greedy selection is not based on well-trained embeddings. In other words, the diversification of MMR is added after model training, while the diversification in paper [63] is added during the model training procedure. This is also the primary distinction between any post-processing and in-processing approaches.

DPP Determinantal Point Process (DPP) is one of the cutting-edge post-processing methods for diversity enhancement in search and recommendation. First introduced by Macchi [85] with the name “fermion process”, DPP was originally used to precisely describe the repulsion and diversity phenomenon for fermion systems in thermal equilibrium. Recently, it has been applied in search and recommendation for enhancing diversity [30].

Prior to the application of DPP in the recommendation, most diversity approaches, such as the basic version of MMR [16], compute the similarity between items in a pairwise way and avoid recommending redundant items to improve diversity. However, these methods are sub-optimal since the pair-wise dissimilarities may not capture complex similarity relationships within the whole list, also the relevance and diversity are captured separately [30]. Thanks to DPP’s outstanding ability to capture the global correlations among data with an elegant probabilistic model [96], DPP-based methods directly model the dissimilarities among items in a set-wise way using a unified model.

The idea of DPP can be demonstrated as follows. A point process \mathcal{P} on a set \mathcal{D} (e.g., a set of $|\mathcal{D}|$ items) is a probability distribution on the powerset of \mathcal{D} (the set of all subsets of \mathcal{D}). That is, $\forall \mathcal{C} \subseteq \mathcal{D}$, \mathcal{P} assigns some probability $p(\mathcal{C})$, and $\sum_{\mathcal{C} \subseteq \mathcal{D}} p(\mathcal{C}) = 1$. Although a DPP defines a probability distribution over an exponential number of sets, it can be compactly parameterized by a single positive semi-definite (PSD) matrix $\mathbf{L} \in \mathbb{R}^{|\mathcal{D}| \times |\mathcal{D}|}$ [97]. The probability of a subset \mathcal{C} represented by a DPP can be written as:

$$p(\mathcal{C}) \propto \det(\mathbf{L}_{\mathcal{C}}). \quad (34)$$

where $\mathbf{L}_{\mathcal{C}} = [\mathbf{L}_{ij}]_{d_i, d_j \in \mathcal{C}}$. However, it is still unclear how the determinant unifies the relevance and diversity. To show this, we offer two views. The first comprehension is based on the geometric meaning of determinant. Since \mathbf{L} is PSD, we can find a matrix \mathbf{B} such that $\mathbf{L} = \mathbf{B}^T \mathbf{B}$. Then we have $\det(\mathbf{L}_{\mathcal{C}}) = \text{Vol}^2(\{\mathbf{B}_i\}_{i \in \mathcal{C}})$, which can be represented as the squared volume of the parallelepiped spanned by the columns of \mathbf{B} corresponding to elements in \mathcal{C} . The volume here is determined by two factors: the magnitude of column

vectors and the orthogonality among them. If we treat the columns of the matrix \mathbf{B} as item embeddings, then it is clear to see that a larger magnitude of the vectors (*higher relevance*) and stronger orthogonality among them (*higher dissimilarity*) lead to a higher volume (*higher determinant*). Thus, the determinant of a matrix unifies both relevance and diversity. The second comprehension is based on a simple case where the matrix is 2×2 : $\mathbf{L}_{\mathcal{C}} = \begin{bmatrix} l_{11} & l_{21} \\ l_{12} & l_{22} \end{bmatrix}$, whose determinant is $\det(\mathbf{L}_{\mathcal{C}}) = l_{11} \cdot l_{22} - l_{21} \cdot l_{12}$. Assuming the diagonal entries indicate the relevance of items and the off-diagonal entries indicate the similarity among items, then the determinant can be represented as relevance minus similarity. Although this comprehension is for a 2-dimensional case, a similar intuition holds for higher dimensions. Based on the above comprehension and intuition, Chen et al. [30] construct user-specific \mathbf{L} as $\mathbf{L} = \text{diag}(\boldsymbol{\xi}) \cdot \mathbf{S} \cdot \text{diag}(\boldsymbol{\xi})$, where $\text{diag}(\boldsymbol{\xi})$ is a diagonal matrix whose diagonal entries are the relevance scores between items to the user, and the (i, j) -th element of \mathbf{S} is the similarity score between the i^{th} item and the j^{th} item. Thus, Eq. 34 can also be written as:

$$p(\mathcal{C}) \propto \det(\mathbf{L}_{\mathcal{C}}) = \left(\prod_{i \in \mathcal{C}} \xi_i^2 \right) \cdot \det(\mathbf{S}_{\mathcal{C}}). \quad (35)$$

Finally, to obtain the diversified top- K recommendation for user u given the whole item set \mathcal{D} , we first construct a user-specific matrix \mathbf{L} as aforementioned. Then the task can be formulated as follows:

$$\text{set}(\sigma_u) = \underset{\mathcal{C} \subseteq (\mathcal{D} \setminus \mathcal{D}_u), |\mathcal{C}|=K}{\operatorname{argmax}} \det(\mathbf{L}_{\mathcal{C}}). \quad (36)$$

However, directly solving this task is expensive. Approximate solutions to Eq. 36 can be obtained by several algorithms, among which the greedy solution was previously considered the fastest one. Initializing σ_u as empty, an item d that maximizes the following equation is added to σ_u iteratively. Specifically, the selection of the k^{th} item for user u can be described as follows:

$$d^k = \underset{d \in (\mathcal{D} \setminus \mathcal{D}_u) \setminus \text{set}(\sigma_u^{1:k-1})}{\operatorname{argmax}} \left(\det(\mathbf{L}_{\text{set}(\sigma_u^{1:k-1}) \cup \{d\}}) - \det(\mathbf{L}_{\text{set}(\sigma_u^{1:k-1})}) \right). \quad (37)$$

Based on the strength of DPP, Gong et al. [38] propose a diversity-aware Web APIs recommendation methodology for choosing diverse and suitable APIs for mashup creation. The APIs recommendation issue for mashup creation is specifically treated as a graph search problem that seeks the smallest group of Steiner trees in an API correlation graph. Their method innovatively employs determinantal point processes to diversify the recommended results.

6.3.2 Refinement-based Methods

Unlike greedy-based methods that iteratively select items to form the entire ranking list, refinement-based methods adjust positions or replace items in existing ranking lists. Typically, using refinement-based methods, items are initially ranked using relevance metrics and subsequently refined by introducing diversity metrics.

Several earlier works follow this line of approach. For instance, Ziegler et al. [20] construct two ranking lists for retrieving the diversified top-K items: σ_{rel} and σ_{div} , where the first list is constructed merely based on the relevance

score, while the second one is constructed based on the diversity score of each item in the whole candidate sets. Both lists rank the items in descending order based on scores. For achieving a single diversified ranking list, the authors merge the two lists using a scaling factor to trade-off how much to rely on the rankings in σ_{rel} or σ_{div} . A similar strategy is used by Yu et al. [86]. Starting from one ranking list with the K highest scoring items, the authors swap the item that contributes the least to the diversity of the entire set with the next highest scoring item from the remaining. They set a threshold for the relevance when replacing the items in order to avoid a dramatic drop in the overall relevance.

7 ONLINE APPROACHES FOR DIVERSITY

So far, we have reviewed the offline approaches for enhancing diversity in search and recommendation. These methods generally train the model in an offline manner using the existing data with ground-truth labels. However, in some situations, these labeled data are insufficient or unavailable, especially in the recommendation scenario. For instance, one of the most well-known challenges is the “cold-start” problem where new users join the system. To resolve these problems, one effective way is to use online approaches where the systems first display item lists to users, gather user feedback, and then update the model for the next turn. Based on whether the user preference change, we further divide them as (i) *bandit strategies* (i.e., invariant user preference) and (ii) general *reinforcement learning* (i.e., dynamic user preference). In this section, we review how to achieve diversity in these approaches.

7.1 Bandit Strategies

As one of the simplest examples of reinforcement learning (RL), the bandit problem was first introduced by Thompson [98] in 1933. The most classical bandit problem is known as the multi-armed bandit (MAB), whose name comes from imagining a gambler at a row of slot machines, who has to decide how to play these machines to gain as much money as possible in a time horizon [99]. A bandit problem can be generally defined as a sequential game between an agent and an environment [100]. The game is played over T rounds (i.e., the time horizon), while in each round $t \in [T]$, the agent first chooses an action A_t from a given set \mathcal{A} , and the environment then reveals a reward $r_t \in R$. The goal of the agent is to maximize the T -step cumulative reward or, equivalently, minimize the T -step cumulative regret. Here, the cumulative regret is defined as the expected difference between the reward sum associated with an optimal strategy and the sum of the collected rewards $\rho = T \cdot \mu^* - \sum_{t=1}^T r_t$, where μ^* is the maximal reward mean associated with the optimal strategy.

It is intuitive to model online search and recommendation as an MAB, where the algorithm is the agent, items are arms, displaying an item is selecting the corresponding arm, and user feedback is the reward. However, MAB does not use state information, or context (i.e., user and item features), which can limit performance, particularly in recommendation where personalization is key. To address this, most works use an MAB extension called Contextual MAB (CMAB) for online search and recommendation problems.

Abundant work shows that CMAB typically outperforms MAB in the relevance of output lists.

Due to the simplicity of implementation and capability of making real-time decisions, recent research also aims to incorporate diversity in bandit algorithms for search and recommendation. There are generally two ways to enhance diversity in these methods: either to treat diversity as part of the scores of each arm or to design a different bandit architecture that can lead to a diversified result. We review both of these two ideas in the following paragraphs.

7.1.1 Diversity as Score

Most works interpret diversity as part of the score on each arm (item) in the bandit algorithms for search and recommendation. For instance, Li et al. [88] formulate the diversified retrieval of top- K items as a bandit problem with cascading user behavior, where a user browses the displayed list from top to bottom, clicks the first attractive item, and stops browsing the rest. If the user clicks an item, the reward is 1, otherwise 0. Then the objective is to minimize the following T -step cumulative regret:

$$R(T) = \sum_{t=1}^T \mathbb{E}[r(\sigma^*, \alpha_t) - r(\sigma^t, \alpha_t)]. \quad (38)$$

Here, $r(\cdot)$ is the binary reward from the user feedback. σ^t is the displayed ranking list at time step t , while σ^* is the optimal ranking list, with constraint that $|\sigma^t| = |\sigma^*| = K$. α_t is a vector of length K , indicating the *attraction* of each arm (item) in the ranking list at time step t , where is how diversity comes in. Specifically, the authors define the *attraction* as a combination of relevance and diversity, following a very similar way to Eq. 28 in Section 6.2.2. Again, all the definitions of diversity are applicable, while both paper [88] and [89] use the gain on coverage of subtopics (S-Coverage) of adding item d_i as the attraction score of which from the diversity component. Several other works choose different ways to define the diversity score. For instance, Qin et al. [90] use the entropy regularizer, while Wang et al. [91] propose three separate solutions, borrowing from MMR [16], entropy regularizer [90], and temporal user-based switching [101].

7.1.2 Diversity as Architecture

Rather than merely treating diversity as part of the score, Parapar and Radlinski [92] design a different bandit architecture for enhancing diversity. Different from prior works that interpret each arm as an individual item, the authors first make each arm represent a unique item category, and further consider retrieving different items under each category. Such a two-stage design can not only guarantee the items are diverse (i.e., satisfy the distance-based metrics), but also guarantee different categories are covered as much as possible (i.e., satisfy coverage-based metrics). In such a way, the algorithm can be efficiently used to construct user profiles with diverse preference elicitation.

All the works above lie in the recommendation scenario, where the personalization is at the core. However, the output of a conventional web search is typically static, so it is more concerned with satisfying a population of users as opposed to each individual. Following this line, Radlinski et al. [93] propose to learn diverse rankings in web search

systems through MAB. Their proposed approach, Ranked Bandits Algorithm (RBA), runs an MAB instance MAB_i for each rank i (i.e., $1 \leq i \leq K$), where the arm of each MAB indicates a unique item. When user u_t arrives at time t , each MAB_i sequentially and independently decides which item to select at the rank i for displaying to u_t . Assuming u_t follows a cascading browsing behavior (i.e., click at most one relevant item in the list), if u_t clicks on an item actually selected by an MAB instance, the reward for the arm corresponding to that item for the MAB at that rank is 1. The reward for the arms corresponding to all other items is 0. As such, each MAB can update the value of each item iteratively through multiple rounds. Although RBA shows effectiveness both empirically and theoretically, it is worth noting that it is hard to be extended to non-binary payoffs.

7.2 Reinforcement Learning

Although bandit strategies show efficiency and effectiveness in online search and recommendation, there exist several obvious limitation in them. Firstly, bandit algorithms has only one state with several actions that lead back to the same state. In other words, they assume that user preference will always remain the same, which does not hold in most real-world scenarios. Secondly, bandit algorithms only care about the immediate reward, while the long-term reward is still significant to real-world users. To address these, most research naturally brings in the reinforcement learning (RL) framework to model the problem, where the state can be affected by the action of agents and the long-term reward is also captured during recommendation.

In the RL setting, diversity has been promoted by employing efficient exploration-exploitation strategies. Zheng et al. [94] first use a Deep Q-Learning Network (DQN) [102] to capture the long-term award of users' actions. As for the diversity, they adopt a Dueling Bandit Gradient Descent (DBGD) [103, 104, 105] algorithm to do exploration in the DQN framework. Specifically, during their exploration strategy, the agent aims to generate a recommendation list σ using the current network Q and another list σ' using an explore network Q' , which shares the same architecture as Q with a small disturbance added on the parameters of Q . Then the authors conduct a probabilistic interleave [103] to generate the merged recommendation list based on L and L' for obtaining a diversified ranking list. Other researchers such as Stamenkovic et al. [35] first define and presents the next item recommendation objective as a Multi-objective MDP problem. Thereafter, they propose a Scalarized Multi-objective Reinforcement Learning (SMORL) model, which works as a regularizer, incorporating desired properties into the recommendation model to balance the relevance, diversity, and novelty of recommendation.

8 APPLICABILITY OF DIVERSITY METRICS AND APPROACHES

We now delve into the applicability of diversity metrics and approaches across various recommendation models.

Diversity metrics are largely model-agnostic, offering a versatile toolkit for evaluating across a wide array of models. However, there are exceptions where the effective application of certain metrics hinges on the availability of

specific types of data. For instance, coverage-based metrics require knowledge of item categories, whereas distance-based metrics rely on the availability of item embeddings, illustrating that while metrics are broadly applicable, their utility may vary depending on certain specifics.

When it comes to diversity approaches, our discussion is primarily centered on offline approaches due to their prevalent use in current research and applications. These approaches can be broadly categorized into pre-processing, in-processing, and post-processing strategies, each with its unique considerations and potential exceptions regarding their integration with different recommendation models. Pre-processing strategies, such as defining user types, establishing sampling strategies, and setting ground truth labels, offer broad applicability but vary in necessity depending on the model's loss function. For instance, pair-wise loss models [49] may benefit more from sampling strategies than point-wise loss models, yet all models can leverage user types and ground truth labels to enhance recommendation precision. In-processing approaches, characterized by their ability to treat diversity either as regularization or directly within the scoring function, show flexibility across models. Although intuitively aligned with list-wise loss models due to their holistic assessment of recommendation lists, these approaches can also be adapted for pair-wise and point-wise models, demonstrating the potential for wide applicability through creative adaptations like constructing lists from pair-wise or point-wise scores to apply diversity constraints [33, 63]. Post-processing strategies stand out for their universal compatibility, enabling the integration of diversity enhancements after model training, thereby maintaining their efficacy across all types of models.

The emergence of Large Language Models (LLMs) introduces new considerations for the application of diversity approaches. LLMs, with their advanced natural language processing capabilities, present unique challenges and opportunities for integrating diversity metrics and approaches. Diversity metrics are still applicable, while the direct application of diversity approaches to LLM-based recommendation systems may not be that straightforward. For instance, applying in-processing approaches to models utilizing LLMs for recommendations via prompt engineering is challenging, due to their unique architectures and data representation [106, 107]. However, when LLMs are used to augment existing recommendation frameworks by enhancing data quality or providing auxiliary information [108, 109], the potential for applying diversity approaches remains viable.

In summary, while diversity metrics and approaches generally maintain a high degree of model-agnosticism, enabling their application across a spectrum of models, certain exceptions and considerations must be acknowledged. This includes the unique challenges posed by the integration of LLMs into the recommendation landscape, underscoring the need for adaptive and flexible strategies to ensure the effective incorporation of diversity considerations in modern search and recommender systems.

9 OPENNESS AND FUTURE DIRECTIONS

Researchers have realized the importance of improving diversity in retrieval systems and have started the exploration. However, we argue that there still exists openness in this

area. In this section, we discuss a number of open challenges and point out some future opportunities in an effort to encourage additional research in this field.

9.1 Time Dependency

Existing research on diversity-aware retrieval systems focuses primarily on a single time point without taking a continuous time span into account. In real-world systems, however, time plays an important role in the study of user behaviors and intentions, as humans may require varying degrees of diversity at different stages of their interaction with the system. We argue that an intriguing future research direction is to investigate how to ensure personalized and time-dependent diversity in a continuous learning setting in which data arrive in a time-series fashion. For instance, when a new user first joins a system, it is reasonable for the algorithm to display more diverse results in order to help the user better explore her interests. As more data is collected about the user’s interaction with the system, the algorithm should be able to adjust itself to adaptively balance relevance and diversity in order to not only provide items that the user likes based on the user’s past preferences, but also show serendipity to the user at some point in order to attract and retain the user.

9.2 Direct Optimization of Metrics

One of the challenges in enhancing the diversity of search results in retrieval systems is that some metrics are difficult to optimize directly. Although methods have been proposed to make some metrics differentiable (e.g., α -nDCG, Gini Index) [110, 111], most metrics, such as coverage-based metrics and SD Index, are difficult to optimize directly. This hinders the capacity of in-processing methods to achieve trade-offs between diversity and other metrics. Exploring a more general method for differentiating these metrics for end-to-end training could be an intriguing line of research.

9.3 Diversity in Explainability

While much of the diversity-aware research on search and recommendation concentrates on presenting a diverse item list to users, diversity can also pertain to other dimensions like explainability, equally important for user retention and satisfaction. For example, it is not ideal if a recommender system always explains to a user as “*based on your previous history*” or “*similar users also like...*”. This research direction has received scant attention, with only a few works exploring this area [28, 44]. We think it is intriguing to explore what user and item features cause varying degrees of diversity in output lists, as this understanding can guide the generation of diverse user explanations.

9.4 Multi-Stakeholder Trade-offs

Managing multi-stakeholder interests in search and recommendation systems presents a complex challenge, involving balancing the needs and preferences of users, content creators, platform owners, society, etc [112]. Currently, there is a dearth of exploration in how to simultaneously ensure relevance, diversity, and fairness among these entities. For example, while a user might desire a highly personalized and diverse set of recommendations, content creators could desire a broad distribution of their content to reach a

wider audience. Similarly, platform owners might want to maximize user engagement and the time spent on their platform while maintaining a fair playing field for all content creators. Current literature on diversity in search and recommendation, however, predominantly centers on users or content creators, rather than embracing a system-wide perspective. Consequently, this has resulted in a scarcity of comprehensive reviews on diversity that incorporate these broader stakeholder perspectives. We observe that there exist strong relations among different stakeholders, for instance, promoting diversity at the individual level (to meet user satisfaction) and at the system level (to cater to content creators) can be seen as enhancing the overall diversity of the system, which in turn satisfies platform owners. Additionally, when including “society” as a stakeholder, the social welfare metrics discussed in Sec. 5.1.3, such as the Simpson’s Diversity Index and the Gini Index, become crucial targets for optimization. Future research in this area could investigate mechanisms to navigate these trade-offs and model the dynamics between multiple stakeholders, providing a more holistic approach to designing diversity-aware retrieval systems.

9.5 Diversity in Multi-Modal Recommendation

Multi-modal recommendation systems, such as those integrating text, images, and audio, pose unique challenges and opportunities for diversity. So far, most diversity-aware systems have focused on single-mode recommendation, such as text-only or image-only recommendations. However, as our digital world becomes increasingly multi-modal, the concept of diversity must expand to consider how different modalities contribute to or detract from diversity. This can be a complex task, given the distinct characteristics and inherent diversity within and across different modalities. For instance, diversity in text might relate to content topic or writing style, while diversity in images might be related to color, style, or content theme. In audio recommendations, diversity could be associated with various factors like genre, artist, or mood. Identifying how to effectively measure and optimize diversity in a multi-modal context is an open challenge that requires further investigation. Future research in this area can bring new insights into how different modalities interact and how they can be combined to create more diverse and enriching user experiences.

9.6 Relation to Other Metrics

Diversity cannot be viewed in isolation; it is interconnected with other important metrics. In our survey, we touch upon how diversity relates to relevance and novelty. Another critical metric that is often discussed in conjunction with diversity is fairness [113, 114]. Enhancing diversity can align with efforts to increase fairness in recommendations, contributing to a more balanced and equitable distribution of content across users, which may further benefit the entire society. Exploring the trade-offs between diversity and fairness metrics presents a compelling future direction.

Moreover, with the rise of LLMs, the evaluation of diversity in natural language generation has garnered more interest. For example, Tevet and Berant [115] provide a summary of diversity metrics in this domain and assess their

effectiveness. There is potential for overlap between diversity measures in natural language generation and those in search and recommendation systems, particularly as the use of LLMs in search and information retrieval becomes more prevalent. Adopting these metrics in search and recommendation contexts could offer a richer evaluation framework, allowing for insights from diverse angles.

10 CONCLUSION

In this survey, we introduce the foundations, definitions, metrics, and approaches of diversity in retrieval systems from the perspective of search and recommendation. We begin the survey with a introduction of why diversity is important in retrieval systems for the benefit of multiple stakeholders. To help better understand the diversity concepts, we summarize the different diversity concerns in search and recommendation, highlighting their connection and distinctions. For the main body of the survey, we provide a unified taxonomy to classify the metrics and approaches of diversification in both search and recommendation. To close the survey, we discuss the open research questions of diversity-aware research in retrieval systems in the hopes of inspiring future innovations and encouraging the deployment of diversity in real-world systems.

REFERENCES

- [1] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, "BPR: bayesian personalized ranking from implicit feedback," in *UAI*, 2009.
- [2] B. M. Sarwar, G. Karypis, J. A. Konstan, and J. Riedl, "Item-based collaborative filtering recommendation algorithms," in *WWW*, 2001.
- [3] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, and M. Wang, "Lightgcn: Simplifying and powering graph convolution network for recommendation," in *SIGIR*, 2020.
- [4] E. Pariser, *The filter bubble: How the new personalized web is changing what we read and how we think*, 2011.
- [5] S. Rosen, "The economics of superstars," *The American economic review*, vol. 71, no. 5, pp. 845–858, 1981.
- [6] Q. Wu, Y. Liu, C. Miao, Y. Zhao, L. Guan, and H. Tang, "Recent advances in diversified recommendation," *CoRR*, vol. abs/1905.06589, 2019.
- [7] J. Sun, W. Guo, D. Zhang, Y. Zhang, F. Regol, Y. Hu, H. Guo, R. Tang, H. Yuan, X. He, and M. Coates, "A framework for recommending accurate and diverse items using bayesian graph convolutional neural networks," in *KDD*, 2020, pp. 2030–2039.
- [8] H. K. Azad and A. Deepak, "Query expansion techniques for information retrieval: A survey," *Inf. Process. Manag.*, vol. 56, no. 5, pp. 1698–1735, 2019.
- [9] L. Azzopardi, "Cognitive biases in search: A review and reflection of cognitive biases in information retrieval," in *CHIIR*, 2021, pp. 27–37.
- [10] W. Huang, B. Liu, and H. Tang, "Privacy protection for recommendation system: a survey," in *Journal of Physics: Conference Series*, vol. 1325, no. 1, 2019, p. 012087.
- [11] J. Chen, H. Dong, X. Wang, F. Feng, M. Wang, and X. He, "Bias and debias in recommender system: A survey and future directions," *CoRR*, vol. abs/2010.03240, 2020.
- [12] C. Faloutsos and D. W. Oard, "A survey of information retrieval and filtering methods," *Tech. Rep.*, 1995.
- [13] S. Zhang, L. Yao, A. Sun, and Y. Tay, "Deep learning based recommender system: A survey and new perspectives," *ACM Comput. Surv.*, vol. 52, no. 1, pp. 5:1–5:38, 2019.
- [14] M. Kunaver and T. Pozrl, "Diversity in recommender systems - A survey," *Knowl. Based Syst.*, 2017.
- [15] J. Chakraborty and V. Verma, "A survey of diversification techniques in recommendation systems," in *SAPIENCE*, 2016, pp. 35–40.
- [16] J. G. Carbonell and J. Goldstein, "The use of mmr, diversity-based reranking for reordering documents and producing summaries," in *SIGIR*, 1998, pp. 335–336.
- [17] C. L. A. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon, "Novelty and diversity in information retrieval evaluation," in *SIGIR*, 2008, pp. 659–666.
- [18] F. Radlinski, P. N. Bennett, B. Carterette, and T. Joachims, "Redundancy, diversity and interdependent document relevance," *SIGIR Forum*, vol. 43, no. 2, pp. 46–52, 2009.
- [19] M. Zhang and N. Hurley, "Avoiding monotony: improving the diversity of recommendation lists," in *RecSys*, 2008, pp. 123–130.
- [20] C. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen, "Improving recommendation lists through topic diversification," in *WWW*, 2005, pp. 22–32.
- [21] P. Cheng, S. Wang, J. Ma, J. Sun, and H. Xiong, "Learning to recommend accurate and diverse items," in *WWW*, 2017, pp. 183–192.
- [22] R. Agrawal, S. Gollapudi, A. Halverson, and S. Jeong, "Diversifying search results," in *WSDM*, 2009, pp. 5–14.
- [23] E. Amigó, D. Spina, and J. Carrillo-de Albornoz, "An axiomatic analysis of diversity evaluation metrics: Introducing the rank-biased utility metric," in *SIGIR*, 2018.
- [24] N. Hurley, "Towards diverse recommendation," *DiveRS@RecSys*, 2011.
- [25] L. Chen, G. Zhang, and E. Zhou, "Fast greedy MAP inference for determinantal point process to improve recommendation diversity," in *NIPS*, 2018, pp. 5627–5638.
- [26] Y. Huang, W. Wang, L. Zhang, and R. Xu, "Sliding spectrum decomposition for diversified recommendation," in *KDD*, 2021, pp. 3041–3049.
- [27] W. Chen, P. Ren, F. Cai, F. Sun, and M. de Rijke, "Improving end-to-end sequential recommendations with intent-aware diversification," in *CIKM*, 2020.
- [28] X. Li, W. Jiang, W. Chen, J. Wu, G. Wang, and K. Li, "Directional and explainable serendipity recommendation," in *WWW*, 2020, pp. 122–132.
- [29] L. Gan, D. Nurbakova, L. Laporte, and S. Calabretto, "Enhancing recommendation diversity using determinantal point processes on knowledge graphs," in

- SIGIR*, 2020, pp. 2001–2004.
- [30] L. Chen, G. Zhang, and H. Zhou, “Improving the diversity of top-n recommendation via determinantal point process,” *CoRR*, vol. abs/1709.05135, 2017.
- [31] Y. Liang, T. Qian, Q. Li, and H. Yin, “Enhancing domain-level and user-level adaptivity in diversified recommendation,” in *SIGIR*, 2021.
- [32] J. Parapar and F. Radlinski, “Diverse user preference elicitation with multi-armed bandits,” in *WSDM*, 2021, pp. 130–138.
- [33] J. Wasilewski and N. Hurley, “Incorporating diversity in a learning to rank recommender system,” in *FLAIRS Conference*, 2016, pp. 572–578.
- [34] S. Vargas, P. Castells, and D. Vallet, “Intent-oriented diversity in recommender systems,” in *SIGIR*, 2011, pp. 1211–1212.
- [35] D. Stamenkovic, A. Karatzoglou, I. Arapakis, X. Xin, and K. Katevas, “Choosing the best of both worlds: Diverse and novel recommendations through multi-objective reinforcement learning,” in *WSDM*, 2022, pp. 957–965.
- [36] W. Chen, P. Ren, F. Cai, F. Sun, and M. De Rijke, “Multi-interest diversification for end-to-end sequential recommendation,” *TOIS*, vol. 40, no. 1, pp. 1–30, 2021.
- [37] J. Han and H. Yamana, “Geographic diversification of recommended pois in frequently visited areas,” *TOIS*, vol. 38, no. 1, pp. 1–39, 2019.
- [38] W. Gong, X. Zhang, Y. Chen, Q. He, A. Beheshti, X. Xu, C. Yan, and L. Qi, “DAWAR: diversity-aware web apis recommendation for mashup creation based on correlation graph,” in *SIGIR*, 2022, pp. 395–404.
- [39] K. Tsukuda and M. Goto, “Dualdiv: diversifying items and explanation styles in explainable hybrid recommendation,” in *RecSys*, 2019, pp. 398–402.
- [40] J. R. Haritsa, “The KNDN problem: A quest for unity in diversity,” *IEEE Data Eng. Bull.*, vol. 32, no. 4, pp. 15–22, 2009.
- [41] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. Riedl, “Evaluating collaborative filtering recommender systems,” *TOIS*, vol. 22, no. 1, pp. 5–53, 2004.
- [42] M. Ge, C. Delgado-Battenfeld, and D. Jannach, “Beyond accuracy: evaluating recommender systems by coverage and serendipity,” in *RecSys*, 2010, pp. 257–260.
- [43] B. Paudel, T. Haas, and A. Bernstein, “Fewer flops at the top: Accuracy, diversity, and regularization in two-class collaborative filtering,” in *RecSys*, 2017, pp. 215–223.
- [44] G. Balloccu, L. Boratto, G. Fenu, and M. Marras, “Post processing recommender systems with knowledge graphs for recency, popularity, and diversity of explanations,” in *SIGIR*, 2022, pp. 646–656.
- [45] G. Adomavicius and Y. Kwon, “Improving aggregate recommendation diversity using ranking-based techniques,” *TKDE*, vol. 24, no. 5, pp. 896–911, 2012.
- [46] K. Raman, P. N. Bennett, and K. Collins-Thompson, “Understanding intrinsic diversity in web search: Improving whole-session relevance,” *TOIS*, vol. 32, no. 4, pp. 1–45, 2014.
- [47] M. Wilhelm, A. Ramanathan, A. Bonomo, S. Jain, E. H. Chi, and J. Gillenwater, “Practical diversified recommendations on youtube with determinantal point processes,” in *CIKM*, 2018, pp. 2165–2173.
- [48] H. Kwon, J. Han, and K. Han, “ART (attractive recommendation tailor): How the diversity of product recommendations affects customer purchase preference in fashion industry?” in *CIKM*, 2020, pp. 2573–2580.
- [49] Y. Zheng, C. Gao, L. Chen, D. Jin, and Y. Li, “DGCN: diversified recommendation with graph convolutional networks,” in *WWW*, 2021, pp. 401–412.
- [50] J. Zhou, E. Agichtein, and S. Kallumadi, “Diversifying multi-aspect search results using simpson’s diversity index,” in *CIKM*, 2020.
- [51] Y. He, H. Zou, H. Yu, Q. Wang, and S. Gao, “Diversity-aware recommendation by user interest domain coverage maximization,” in *ICDM*, 2019, pp. 1084–1089.
- [52] X. Yin, J. X. Huang, Z. Li, and X. Zhou, “A survival modeling approach to biomedical search result diversification using wikipedia,” *TKDE*, vol. 25, no. 6, pp. 1201–1212, 2013.
- [53] R. Li and J. X. Yu, “Scalable diversified ranking on large graphs,” *TKDE*, vol. 25, no. 9, pp. 2133–2146, 2013.
- [54] C. Zhai, W. W. Cohen, and J. D. Lafferty, “Beyond independent relevance: methods and evaluation metrics for subtopic retrieval,” in *SIGIR*, 2003, pp. 10–17.
- [55] X. Qin, Z. Dou, and J. Wen, “Diversifying search results using self-attention network,” in *CIKM*, 2020.
- [56] S. Liang, F. Cai, Z. Ren, and M. de Rijke, “Efficient structured learning for personalized diversification,” *TKDE*, vol. 28, no. 11, pp. 2958–2973, 2016.
- [57] S. Liang, E. Yilmaz, H. Shen, M. D. Rijke, and W. B. Croft, “Search result diversification in short text streams,” *TOIS*, vol. 36, no. 1, pp. 1–35, 2017.
- [58] E. H. Simpson, “Measurement of diversity,” *nature*, vol. 163, no. 4148, pp. 688–688, 1949.
- [59] A. Antikacioglu and R. Ravi, “Post processing recommender systems for diversity,” in *KDD*, 2017, pp. 707–716.
- [60] J. Sanz-Cruzado and P. Castells, “Enhancing structural diversity in social networks by recommending weak ties,” in *RecSys*, 2018, pp. 233–241.
- [61] C. Gini, “Variabilità e mutabilità,” 1912.
- [62] J. Parapar and F. Radlinski, “Towards unified metrics for accuracy and diversity for recommender systems,” in *RecSys*, 2021, pp. 75–84.
- [63] S. Li, Y. Zhou, D. Zhang, Y. Zhang, and X. Lan, “Learning to diversify recommendations based on matrix factorization,” in *DASC/PiCom/DataCom/CyberSciTech*, 2017, pp. 68–74.
- [64] R. L. T. Santos, C. Macdonald, and I. Ounis, “Exploiting query reformulations for web search result diversification,” in *WWW*, 2010, pp. 881–890.
- [65] H. Yu, “Optimize what you evaluate with: Search result diversification based on metric optimization,” in *AAAI*, 2022, pp. 10399–10407.
- [66] Z. Jiang, Z. Dou, W. X. Zhao, J. Nie, M. Yue, and J. Wen, “Supervised search result diversification via subtopic attention,” *TKDE*, vol. 30, no. 10, pp. 1971–1984, 2018.
- [67] F. Cai, R. Reinanda, and M. D. Rijke, “Diversifying

- query auto-completion," *TOIS*, vol. 34, no. 4, pp. 1–33, 2016.
- [68] Y. Yang, A. Lad, N. Lao, A. Harpale, B. Kisiel, and M. Rogati, "Utility-based information distillation over temporally sequenced documents," in *SIGIR*, 2007, pp. 31–38.
- [69] Y. Yang and A. Lad, "Modeling expected utility of multi-session information distillation," in *ICTIR*, 2009, pp. 164–175.
- [70] T. Sakai and Z. Zeng, "Retrieval evaluation measures that agree with users' serp preferences: Traditional, preference-based, and diversity measures," *TOIS*, vol. 39, no. 2, pp. 1–35, 2020.
- [71] T. Sakai, "Evaluation with informational and navigational intents," in *WWW*, 2012, pp. 499–508.
- [72] T. Sakai and R. Song, "Evaluating diversified search results using per-intent graded relevance," in *SIGIR*, 2011, pp. 1043–1052.
- [73] T. Sakai, N. Craswell, R. Song, S. Robertson, Z. Dou, and C. Lin, "Simple evaluation metrics for diversified search results," in *EVA*, 2010, pp. 42–50.
- [74] X. Wang, J. Wen, Z. Dou, T. Sakai, and R. Zhang, "Search result diversity evaluation based on intent hierarchies," *TKDE*, vol. 30, no. 1, pp. 156–169, 2018.
- [75] C. Yu, L. V. S. Lakshmanan, and S. Amer-Yahia, "It takes variety to make a world: diversification in recommender systems," in *EDBT*, 2009, pp. 368–378.
- [76] J. C. Gower, "A general coefficient of similarity and some of its properties," *Biometrics*, vol. 27, no. 4, pp. 857–871, 1971.
- [77] A. Moffat and J. Zobel, "Rank-biased precision for measurement of retrieval effectiveness," *TOIS*, vol. 27, no. 1, pp. 2:1–2:27, 2008.
- [78] E. Amigó, J. Gonzalo, and F. Verdejo, "A general evaluation measure for document organization tasks," in *SIGIR*, 2013, pp. 643–652.
- [79] E. Amigó, D. Spina, and J. C. de Albornoz, "An axiomatic analysis of diversity evaluation metrics: Introducing the rank-biased utility metric," in *SIGIR*, 2018, pp. 625–634.
- [80] B. Carterette, "An analysis of np-completeness in novelty and diversity ranking," *Information Retrieval*, vol. 14, pp. 89–106, 2011.
- [81] C. L. Clarke, M. Kolla, and O. Vechtomova, "An effectiveness measure for ambiguous and underspecified queries," in *ICTIR*, 2009, pp. 188–199.
- [82] S. Maropaki, S. Chester, C. Doulkeridis, and K. Nørvåg, "Diversifying top-k point-of-interest queries via collective social reach," in *CIKM*, 2020.
- [83] L. Chen and H. Shi, "Dexdeepfm: Ensemble diversity enhanced extreme deep factorization machine model," *TKDD*, vol. 16, no. 5, pp. 1–17, 2022.
- [84] Y. Gu, G. Liu, J. Qi, H. Xu, G. Yu, and R. Zhang, "The moving K diversified nearest neighbor query," *TKDE*, vol. 28, no. 10, pp. 2778–2792, 2016.
- [85] O. Macchi, "The coincidence approach to stochastic point processes," *Advances in Applied Probability*, vol. 7, no. 1, p. 83–122, 1975.
- [86] C. Yu, L. V. S. Lakshmanan, and S. Amer-Yahia, "It takes variety to make a world: diversification in recommender systems," in *EDBT*, 2009, pp. 368–378.
- [87] C. Zhang, H. Liang, and K. Wang, "Trip recommendation meets real-world constraints: Poi availability, diversity, and traveling time uncertainty," *TOIS*, vol. 35, no. 1, pp. 1–28, 2016.
- [88] C. Li, H. Feng, and M. d. Rijke, "Cascading hybrid bandits: Online learning to rank for relevance and diversity," in *RecSys*, 2020.
- [89] Q. Ding, Y. Liu, C. Miao, F. Cheng, and H. Tang, "A hybrid bandit framework for diversified recommendation," in *AAAI*, 2021, pp. 4036–4044.
- [90] L. Qin, S. Chen, and X. Zhu, "Contextual combinatorial bandit and its application on diversified online recommendation," in *SDM*, 2014, pp. 461–469.
- [91] L. Wang, C. Wang, K. Wang, and X. He, "Biucb: A contextual bandit algorithm for cold-start and diversified recommendation," in *ICBK*, 2017, pp. 248–253.
- [92] J. Parapar and F. Radlinski, "Diverse user preference elicitation with multi-armed bandits," in *WSDM*, 2021, pp. 130–138.
- [93] F. Radlinski, R. Kleinberg, and T. Joachims, "Learning diverse rankings with multi-armed bandits," in *ICML*, 2008, pp. 784–791.
- [94] G. Zheng, F. Zhang, Z. Zheng, Y. Xiang, N. J. Yuan, X. Xie, and Z. Li, "DRN: A deep reinforcement learning framework for news recommendation," in *WWW*, 2018, pp. 167–176.
- [95] R. Baeza-Yates, "Modern information retrieval," *Addison Wesley google schola*, vol. 2, pp. 127–136, 1999.
- [96] A. Kulesza and B. Taskar, "Determinantal point processes for machine learning," *Found. Trends Mach. Learn.*, vol. 5, no. 2–3, pp. 123–286, 2012.
- [97] A. Borodin, "Determinantal point processes," *arXiv preprint arXiv:0911.1153*, 2009.
- [98] W. R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, vol. 25, no. 3–4, pp. 285–294, 1933.
- [99] R. Weber, "On the Gittins Index for Multiarmed Bandits," *The Annals of Applied Probability*, vol. 2, no. 4, pp. 1024 – 1033, 1992.
- [100] T. Lattimore and C. Szepesvári, *Bandit Algorithms*, 2020.
- [101] N. Lathia, S. Hailes, L. Capra, and X. Amatriain, "Temporal diversity in recommender systems," in *SIGIR*, 2010, pp. 210–217.
- [102] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. A. Riedmiller, A. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nat.*, vol. 518, no. 7540, pp. 529–533, 2015.
- [103] A. Grotov and M. de Rijke, "Online learning to rank for information retrieval: SIGIR 2016 tutorial," in *SIGIR*, 2016, pp. 1215–1218.
- [104] K. Hofmann, A. Schuth, S. Whiteson, and M. de Rijke, "Reusing historical interaction data for faster online learning to rank for IR," in *WSDM*, 2013, pp. 183–192.
- [105] Y. Yue and T. Joachims, "Interactively optimizing information retrieval systems as a dueling bandits problem," in *ICML*, 2009, pp. 1201–1208.

- [106] Z. Chen, "Palr: Personalization aware llms for recommendation," *arXiv preprint arXiv:2305.07622*, 2023.
- [107] F. Radlinski, K. Balog, F. Diaz, L. Dixon, and B. Wedin, "On natural language user profiles for transparent and scrutable recommendation," in *SIGIR*, 2022, pp. 2863–2874.
- [108] W. Wei, X. Ren, J. Tang, Q. Wang, L. Su, S. Cheng, J. Wang, D. Yin, and C. Huang, "Llmrec: Large language models with graph augmentation for recommendation," *WSDM*, 2024.
- [109] X. Ren, W. Wei, L. Xia, L. Su, S. Cheng, J. Wang, D. Yin, and C. Huang, "Representation learning with large language models for recommendation," *WWW*, 2024.
- [110] M. Wu, Y. Chang, Z. Zheng, and H. Zha, "Smoothing DCG for learning to rank: a novel approach using smoothed hinge functions," in *CIKM*, 2009, pp. 1923–1926.
- [111] V. Do and N. Usunier, "Optimizing generalized gini indices for fairness in rankings," in *SIGIR*, 2022, pp. 737–747.
- [112] H. Abdollahpouri, G. Adomavicius, R. Burke, I. Guy, D. Jannach, T. Kamishima, J. Krasnodebski, and L. Pizzato, "Multistakeholder recommendation: Survey and research directions," *User Modeling and User-Adapted Interaction*, vol. 30, pp. 127–158, 2020.
- [113] R. Gao and C. Shah, "Addressing bias and fairness in search systems," in *SIGIR*, 2021, pp. 2643–2646.
- [114] S. Verma, R. Gao, and C. Shah, "Facets of fairness in search and recommendation," in *Bias and Social Aspects in Search and Recommendation: First International Workshop, BIAS 2020, Lisbon, Portugal, April 14, Proceedings 1*. Springer, 2020, pp. 1–11.
- [115] G. Tevet and J. Berant, "Evaluating the evaluation of diversity in natural language generation," *arXiv preprint arXiv:2004.02990*, 2020.



Haolun Wu is a Ph.D. candidate in Computer Science at Mila - Quebec AI Institute and McGill University. His research interests include information retrieval, recommender systems, knowledge modeling, LLMs, etc. He has published papers on ICLR, SIGIR, AAAI, CIKM, TOIS, ICDE, CHI, etc. He is a student researcher at Google Research and Microsoft Research. He holds the Borealis AI Fellowship 2023-24.



Yansen Zhang is a Ph.D. candidate in Computer Science at City University of Hong Kong. He received his B.S. and M.S. degrees in Software Engineering from NEU and SYSU, China, in 2019 and 2022, respectively. His research interests include diversified recommendation and explainable recommendation. He has published papers on SIGIR and ICONIP.



Chen Ma is currently an Assistant Professor in the Department of Computer Science, City University of Hong Kong. His research interests lie in the theory of data mining and machine learning and their applications in recommender systems, knowledge graphs, social computing, and social good. He received his B.S. and M.S. degrees in Software Engineering from Beijing Institute of Technology in 2013 and 2015, respectively. He received his PhD degree in Computer Science from McGill University.



Fuyuan Lyu is a Ph.D. candidate in Computer Science at McGill University. He obtained his B.S. degree from Shanghai Jiao Tong University with Zhiyuan honor degree. His research interest lies in the intersection between AutoML and IR. He has published papers on NeurIPS, WWW, KDD, ICDE, CIKM, etc. He interned and collaborated with Huawei Noah's Ark Lab and Tencent Financial Technology. He previously worked as a research assistant at Nanyang Technological University and Shanghai Jiao Tong University.



Bowei He is a Ph.D. candidate in Computer Science at City University of Hong Kong. His research focuses on advanced ML (e.g., RL, continual learning, and Auto-ML) and causal inference for data mining and information retrieval. He has published papers on WWW, NeurIPS, ICDE, CVPR, ICCV, CIKM, ICDM, UAI, SDM, ICME, ICASSP, IROS, etc. He interned and collaborated with Didi AI Labs, Huawei Noah's Ark Lab, and Tencent Financial Technology.



Bhaskar Mitra is a Principal Researcher in the Collaborative Intelligence group at Microsoft Research. His research interests lie at the intersections of deep learning, information retrieval, knowledge bases, benchmarking and evaluation, and FATE (Fairness, Accountability, Transparency, and Ethics). He co-created the MS MARCO benchmark and co-organizes the TREC Deep Learning and Tip-of-the-Tongue Tracks. He received his Ph.D. in Computer Science from University College London.



Xue Liu (Fellow, IEEE) is a Professor and a William Dawson Scholar in the School of Computer Science, McGill University. He is a Fellow of the Canadian Academy of Engineering (FCAE) and a Fellow of IEEE (FIEEE). He is an associate member of Mila, and also the Chief Scientist and Co-Director of Samsung AI Center Montreal. He was also the Samuel R. Thompson Associate Professor with the University of Nebraska-Lincoln and HP Labs, Palo Alto, USA. He received the B.S. degree in mathematics and the M.S. degree in automatic control from Tsinghua University in 1996 and 1999, respectively, and the Ph.D. degree in computer science from the University of Illinois at Urbana-Champaign in 2006. He has over 150 published research papers. His areas of interest encompass computer networks and communications, smart grid, real-time and embedded systems, cyber-physical systems, data centers, and applied ML.