Knowl Inf Syst (2017) 51:1–36
DOI 10.1007/s10115-016-0990-4

SURVEY PAPER

A survey of query result diversification

Kaiping Zheng[1] · **Hongzhi Wang**[1] · **Zhixin Qi**[1] · **Jianzhong Li**[1] · **Hong Gao**[1]

Received: 15 March 2015 / Accepted: 8 September 2016 / Published online: 21 September 2016

Abstract Nowadays, in information systems such as web search engines and databases, diversity is becoming increasingly essential and getting more and more attention for improving users' satisfaction. In this sense, query result diversification is of vital importance and well worth researching. Some issues such as the definition of diversification and efficient diverse query processing are more challenging to handle in information systems. Many researchers have focused on various dimensions of diversify problem. In this survey, we aim to provide a thorough review of a wide range of result diversification techniques including various definitions of diversifications, corresponding algorithms, diversification technique specified for some applications including database, search engines, recommendation systems, graphs, time series and data streams as well as result diversification systems. We also propose some open research directions, which are challenging and have not been explored up till now, to improve the quality of query results.

Keywords Diversity · Query processing · Information retrieval

1 Introduction

In the past, web search engines or databases answer queries generally according to the relevance, which is how related a result is to the query. However, recently, people increasingly tend to focus on diversity. That is to say, how different the results returned are, and query result diversification has attracted a lot of attention as it can contribute to enhancing users' satisfaction and improving users' experience. In web search engines and recommendation systems, query result diversification helps counteract the overspecialization problem in which the results retrieved are too homogeneous to meet users' needs [22,94,110].

✉ Hongzhi Wang
wangzh@hit.edu.cn

[1] Harbin Institute of Technology, Harbin, China

Query result diversification is meaningful in several aspects. Firstly, diversification can help avoid the retrieval of overly simplistic results. Therefore, by applying diversification methods, we can achieve a win–win situation. For example, if a user wants to search some information about “computers”, several different topics related to it should be returned, such as the description of computers’ structure and utilities, how and where to buy computers, some academic information about computers and so on.

Secondly, as the users’ queries may be ambiguous at times, it is significant to guess the users’ true intentions in order to offer reasonable and satisfactory results for them. In this sense, diversification should be used since it can describe various conjectures about the query and then, according to these guesses, provide a comprehensive, full-scale final result set. Consider an example. If the user inputs “Harry Potter” as the keyword in a search engine, the final query result set should include books, movies and even the actors for the movie. Only by covering as many as various facts and offering various enough information, can we make users satisfied.

Thirdly, diversification also aims to provide results that can bring new information not previously mentioned. Thus, users can access novel and relevant resources. For instance, when the movie “Harry Potter and the Deathly Hallows: Part 2” comes to market for the first time, users should be offered information about it if they search “Harry Potter”. This is beneficial to both the users in that they can have access to the most fashionable type and the producers in that they are able to better advertise and generalize their new products to people.

1.1 Categorization of diversification methods

Due to the importance, diversification draws attentions of researchers and many approaches have been proposed. For the convenience of the survey, we classify existing approaches. The major differences between the categories are in two dimensions. One is the difference in the definitions of diversity. The other is the difference in the applications.

For the difference in diversity definition, query result diversification is classified into three categories [22], (i) content-based (or similarity-based) diversification, i.e., objects that are dissimilar to each other [24,110], (ii) intent-based (or semantic coverage-based) diversification, i.e., objects that belong to different categories or are related to various topics to respond to users’ ambiguous query requests and to satisfy users’ expected intents [1,22,24], (iii) novelty-based diversification, i.e., objects that contain new information different from the ones previously retrieved in order to improve users’ query satisfaction [16,106].

According to the difference in applications, the approaches could be classified into the diversification for top-k queries, information retrieval, recommendation, time series and so on. Since some approaches have been proposed specially for some applications, we introduce such special approaches in Sect. 5.

1.2 Contributions

As stated above, query result diversification gets more and more attention to improve users’ satisfaction, so it is essential to conclude and summarize the existing works in related literatures. This survey just performs this task. It firstly summarizes several categories of diversification along with their definitions, and some research results in specialized search areas, together with some novel techniques. Then this survey concludes the essences in these works and analyses the contributions of these works from different perspectives. Last but not the least, this survey carries forward the advantages and rectifies the disadvantages of current methods and then proposes the future challenging directions which are worth further exploring.

In this paper, we make the following contributions:

- We discuss several categorizations of diversification methods. Then based on the classic categorization, we present, and compare the definitions of three versions of diversification. Based on these definitions, we analyze, and discuss algorithms corresponding to the definitions, and then make comparisons among them after illustrating the strengths as well as the weaknesses of these algorithms.
- We demonstrate the diversification methods in several specialized search areas including top-k problem, web search, information retrieval, recommendation systems, time series search, graph studies, structured data, skylines, streaming data, keyword search and multi-dimensional search. We also present the situation where diversification is regarded as a spatial concept and enrich the application range of diversification methods.
- We give an introduction about novel techniques, systems together with tools utilized in query result diversification. Therefore, we are able to show that diversification methods can be widely applied in practice to make returned query results more satisfactory to users.
- Finally, we propose three research challenges worth further researching and exploring. Firstly, we may lay more emphases in the implementation of novelty-based diversification. Secondly, we can take the notion "freshness" into consideration in order to better improve users' satisfaction. Thirdly, we establish a methodology to combine these three categories of diversification.

The rest of this paper is structured as follows. Section 2 describes variable definitions for diversification. Content-based and intent-based diversification algorithms are surveyed in Sects. 3 and 4, respectively. Diversification methods on specialized data types and query tasks are illustrated in Sect. 5. Next, Sect. 6 describes various query diversification systems. Miscellaneous topics are discussed in Sect. 7. After this, Sect. 8 puts forward several open research challenges and Sect. 9 concludes this paper.

2 Problem definitions

In this section, we introduce the definition and general settings for the diversification problem with some notations used thorough this paper.

Given a set of data objects X from the universe U and a query q, the general problem of diversification is to maximize the diversity on the results set R over X. We denote all results of q in X as X_q.

For two data objects $o_i, o_j \in X$, their similarity and distance are denoted as $\mathrm{sim}\,(o_i, o_j)$ and $\mathrm{dis}\,(o_i, o_j)$, respectively. The relevance between each object o_i and the query request q can be denoted as $\delta_{\mathrm{sim}}(q, o_i)$. In some settings, relevance and diversity are tradeoff [18]. We denote the factor of the tradeoff to show the importance of relevance and diversity as λ $(0 \leq \lambda \leq 1)$. Sometimes, user information is used. We use P to represent the user set and $\mathrm{Objects}(u)$ to represent the set of all objects which the user u has rated in the past. $\delta_{\mathrm{div}}(o_i, o_j))$ is the diverse of o_i and o_j.

According to different definitions, diversification could be classified into three categories as discussed in Sect. 1.1. The similarity among these categories is that all of them focus on improving the experiences and satisfactory of users. Essentially, these categories are different in their focuses. Content-based diversification emphasizes how different the objects are; intent-based diversification focuses on different facets of a given query; novelty-based diversification lays emphasis on searching for novel information different from obtained information. We introduce the definitions in these categories.

2.1 Content-based diversification

Content-based diversification is also known as similarity-based diversification. It aims to present the dissimilarity between each two objects, which is usually quantized as the distance between each pair of objects. Intuitively, diversity could be achieved by clustering the relevant results according to their similarity and picking results from various clusters. However, such approach is coarse for the diversification for two reasons. The first is that the clustering approach treats the relevance of all results as the same, but they are not the same. The second is that the number of real clusters can hardly be equal to the number of required results. This causes the difficulty in selecting the final results from clusters. Thus, computing a relevant and diverse result set offers a more refined and complex problem.

Definition based on semantic distance In content-based diversification, we often need the computation of content similarity. Semantic distance is used to describe the content difference between two objects. In Gollapudi and Sharma [32], Gollapudi et al. propose a sketching algorithm to compute the pairwise semantic distance on the basis of min-hashing scheme [7, 31]. The algorithm first computes the sketch of each object and then uses Jaccard similarity between sketches to compute the pairwise semantic distance between objects. For instance, if we use "database" as the keyword on http://scholar.google.com, then we can obtain novel database systems recently proposed. If the diversity of results is maintained by a simple sketching algorithm in [32], then the process will involve three steps. (1) The sketches of objects are generated; (2) the pairwise semantic distances are computed; and (3) the semantic distances above some thresholds are controlled using Jaccard similarity coefficient.

For documents, min-hash could be used for such approach. $\mathrm{MH}_h(X)$ is the element in X whose hash value is minimum. That is, $\mathrm{MH}_h(X) = \arg\min_x\{h(x)|x \in B\}$. Then, given k hash functions, denoted as $h_1, h_2, \ldots, h_k$, the sketch of a document d is $S(d) = \{\mathrm{MH}_{h_1}(d), \mathrm{MH}_{h_2}(d), \ldots, \mathrm{MH}_{h_k}(d)\}$. Then, based on the metric-Jaccard similarity coefficient, $\mathrm{sim}(o_i, o_j) = \dfrac{|S(o_i) \bigcap S(o_j)|}{|S(o_i) \bigcup S(o_j)|}$.

Definition based on categorical distance Semantic distance does not work so well in some scenarios. For instance, http://www.apache.org/ and http://www.apache.org/docs intuitively should have similar contents with regard to the web graph. However, with the metric-semantic distance, they actually have very different sketches [32]. In order to solve this problem, taxonomies offer an encoding of distance where the category of a page represents the page's sketch [22,32]. In detail, Gollapudi et al. [32] design a weighted tree distance to measure the similarity between two categories in the taxonomy. Formally, the distance $\mathrm{dis}(u, v) = \sum_{i=1}^{l(u)} \dfrac{1}{2^{e(i-1)}} + \sum_{i=1}^{l(v)} \dfrac{1}{2^{e(i-1)}}$. In this formula, $e \geq 0$ and $l(\cdot)$ denotes the depth of the given node in the taxonomy tree. With such definition, in the example about http://www.apache.org/ and http://www.apache.org/docs, we can draw a conclusion that these two are in similar categories or even the same category, so that they are close to each other in terms of categorical distance. In this way, we safely avoid the problem caused by semantic distance.

Additionally, as an object may belong to multiple categories in practice, the categorical distance cannot be equivalent to the distance between nodes in the taxonomy. With regard to this situation, we generalize categorical distance. Given two objects o_i, o_j and their categorical C_{o_i}, C_{o_j}, the categorical distance between them is computed as follows:

$$\mathrm{dis}_c(o_i, o_j) = \sum_{u \in C_{o_i}, v \in C_{o_j}} \min(C_{o_i}(u), C_{o_j}(v)) \arg\min_v d(u, v)$$

where $C_o(u)$ denotes the confidence of o belonging to category u.

This definition of a weighted tree distance $\mathrm{dis}(u, v)$ reduces to the well-known tree distance when e is set to zero and to the notion of hierarchically separated trees for greater values of e. Thus, nodes corresponding to more general categories (e.g., /Top/Health and /Top/Finance) are more separated than specific categories (e.g., /Top/Health/Geriatrics/Osteoporosis and /Top/Health/Geriatrics/Mental Health). This notion of distance can be extended to the case where a document belongs to multiple categories (with different confidences).

Definition based on context When an object is described with multiple attributes $A = (A_1, A_2, \ldots, A_k)$, i.e., relational tuple or web pages, the diversity could be defined according to the importance of the attributes [94]. In the definition, a diversity ordering is a total order on A denoted by $\prec_r$. That is, if $A_i \prec A_j$, A_i is more important than A_j. For instance, if we intend to purchase a ring in http://www.ebay.com, several attributes are involved including categories, style, metal, ring size, main stone, brand, condition and price (CA, ST, ME, RS, MS, CO, PR for short respectively). Then suppose the priority order is (CA, ST, ME, RS, MS, CO, PR).

For objects with attribute set A and order $\prec$, the similarity function of the lth attribute is described as follows:

$$\mathrm{sim}_l(o_i, o_j) = \begin{cases} 1 & \text{if } o_i \cdot A_l = o_j.A_l \\ 0 & \text{otherwise} \end{cases}$$

A result set R for a query q is diverse with regard to ξ, a prefix of A, if $\sum_{a,b \in S_\xi} \mathrm{sim}_\xi(a, b)$ is minimized among all subset of query results of q.

Definition based on the k-nearest neighbor Haritsa et al. [34] propose a definition of diversity based on the Gower coefficient as a weighted average of respective attribute value differences of the objects. Formally, the process has two phases. Firstly, given two objects o_i and o_j, compute the absolute difference between the attribute values of their corresponding attributes A_i and A_j in diversity-space. Assume that an object has m attributes, then these difference values are sorted in descending order, and denoted as $(\eta_1, \eta_2, \ldots, \eta_m)$. Secondly, we calculate the diversity distance between o_i and o_j as $\mathrm{dis}(o_i, o_j) = \sum_{u=1}^{m} w_u \cdot \eta_u$, where η_u is equal to the uth attribute value difference of o_i and o_j, and w_u is the weight for u.

As stated in [34], given a threshold value ϕ, two objects o_i and o_j are diverse if $\mathrm{dis}(o_i, o_j) > \phi$. Then a set of objects is diverse if every two objects in this set are diverse.

Santos et al. [84] present the “Better Results with Influence Diversification (BRID)” technique. The technique is the basis to the k-diverse nearest neighbor (BRIDk) and to the range diverse (BRIDr) algorithms, which execute k-nearest neighbor and range queries with diversification, showing that the technique can be applied to diversify any type of similarity queries.

Definition based on explanations A formal notion of diversity based on explanations is proposed by Yu et al. [22,104]. With regard to content-based diversification, the explanation of a recommended item is defined as a set of similar objects which the user has highly rated in the past. Formally, the explanation of an object $o_i \in X$ recommended to a user $u \in U$ can be defined as follows: $\mathrm{Expl}(u, o_i) = \{o_j \in X | \mathrm{sim}(o_i, o_j) > \theta \text{ and } o_j \in \mathrm{Objects}(u)\}$, meaning the set of objects is similar to the object o according to a certain criterion and this set of objects has been rated by the user u in the past.

We use the example in http://www.amazon.com. In the explanation-based definition, the explanation of a book is interpreted as a set of similar books which the user has rated high or obviously bought. However, if we try to extend this definition, by considering the explanation of a book as the books bought or highly rated by not only this exact user, but also users who have bought this recommended book, i.e., users who have the same tastes. Then when we check a new book in this web site, we can find "Frequently Bought Together" which offers information mined from users' behavior about the books bought together usually, and also find "Customers Who Bought This Item Also Bought" which provides advice about what to buy through other users' choices with similar tastes in books. Then with two sources of books' explanations, namely "Frequent Bought Together" as well as "Customers Who Bought This Item Also Bought", we can compute the explanation of a book recommended to a certain user. According to a certain user, if two recommended books are similar in their explanations, then we select only one from these two to maintain the diversity of recommendation lists.

Context-based definition in publish/subscribe systems Publish/subscribe systems provide a proactive way of sharing information. As stated in [27], in these systems, a user describes the information needs and interests in subscriptions, and he will be informed whenever other users have posted some information which meets his requirements. Additionally, the subscriptions posted by a certain user are typically of equal importance, and this user will be notified as long as there is a published event satisfying all the user's subscriptions. Therefore, the information and the topics delivered from publishers to subscribers are affected or even controlled by the subscriptions provided by subscribers. However, as for users, they do not only want to get information relevant to their needs, they also desire to obtain as much and diverse information as possible in limited time. In this sense, content-diversification is considered essential and significant in publish/subscribe systems.

In Drosou et al. [27] and Drosou and Pitoura [21], in the context of publish/subscribe systems, a content-based diversification definition is proposed. According to object delivery, three fundamental models for publish/subscribe systems are defined: (i) periodic; (ii) sliding window; (iii) history-based filtering delivery. Then formally in [21], given a number k, a sliding window of length l, a period T, according to the objective function which aims to maximize the average distance of any two points o_i and o_j, i.e., $f(S) = \frac{2}{k(k-1)} \sum_{i=1}^{k} \sum_{j>i}^{k} \text{dis}(o_i, o_j)$, the k most diverse objects are returned to the subscribers from publishers.

A typical example is the blog. In this scenario, a user chooses to pay attention to others' blogs according to the own interests. Then if the chosen bloggers have posted some new passages, this user will be notified. If the content-based diversification definition and its corresponding techniques discussed above are applied in this blog scenario, users' satisfaction will be improved to a great extent, as the users can access a huge amount of information both relevant to their needs and as various as possible.

2.2 Intent-based diversification

Intent-based diversification is for users' ambiguous queries [2,79]. Take the following scenario as an example. A user type the keyword "Harry Potter" in the search box of http://www.bing.com, most returned items are about the book. Perhaps many people are interested in the information listed. However, some users may just want information about the movie.

From the example above, different users may have diverse intents with regard to the same keywords. When it comes to a certain user's information needs according to an ambiguous query, it is not enough to provide only one possible interpretation. Therefore, intent-based diversification (also known as coverage-based diversification) is proposed. When dealing with an ambiguous query request without any further information to disambiguate user's intent, a set of results covering possibly all the different interpretations should be returned.

DIVERSIFY(k) definition A DIVERSIFY(k) problem is defined in [1,22]. Given a query q, a taxonomy with n independent categories $C = \{c_1, c_2, \ldots, c_n\}$, the conditional probability distribution of each category c_i for q is denoted as $P(c_i|q)$, i.e., the probability that c_i is relevant to q. Additionally, the conditional probability distribution of each object o_j being relevant to each category c_i according to the query q is denoted as $Pr(o_j|q, c_i)$, and then given an integer k, the objective is to find k objects which cover as many important interpretations as possible.

Formally, a result set $R \subset X_q$ with $|R| = k$ is obtained to maximize $P(R|q) = \sum_{c_i \in C} P(c_i|q)(1 - \prod_{o_j \in R}(1 - Pr(o_j|q, c_i)))$. This formula aims to get R that maximizes the possibility of each category being covered by at least one object. In this way, we obtain information covering different facets of q.

We use an example to illustrate the procedure of improving intent-based diversification. We use "Harry Potter" as input, which can refer to any book or movie of this series. These two interpretations just respectively lie in two categories *book* and *movie*. When we choose to search results relevant to "Harry Potter" from all departments, we will find that the products are about various books. This query result set is far from satisfactory, as it fails to provide us with information about the *Poster* category. In order to avoid the situations above, information relevant to all the categories of objects should be offered to users. This is the motivation of DIVERSIFY(k).

The weights of interpretation are important for intent-based diversification. Ozdemiray and Altingovde [76] address this problem by estimating the retrieval effectiveness of each aspect.

QL_DIVERSIFY(k) definition Capannini et al. [9] extend DIVERSIFY(k) to QL_DIVERSIFY(k). QL_DIVERSIFY(k) aims to obtain the result set $R \subset X_q$ with $|R| = k$ to maximize the following objective function $P(R|q) = \sum_{c_i \in C} P(c_i|q) \cdot (1 - \prod_{o \in R}(1 - \widetilde{U}(o|X_{c_i})))$, where $\widetilde{U}(r|O_{c_i})$ is the normalized utility of $o \in R$ for q' to describe how good a result o satisfies a user's intent, X_{c_i} is the object belonging to c_i. In this definition, $Pr(x_j|q, c_i)$ is similar to the result's utility $\widetilde{U}(r|O_{q'})$.

When we analyze the example of processing the query with keyword "Harry Potter" on the search engine, it may be found that DIVERSIFY(k) definition and QL_DIVERSIFY(k) definition have much in common. They are almost the same, except that the DIVERSIFY(k) problem focuses more on trying to retrieve enough results on various categories of objects, while the QL_DIVERSIFY(k) problem pays more attention to finding more interpretations, i.e., specializations of a certain query, and then uses these specializations to conduct search operations.

Another feature shared by DIVERSIFY(k) and the QL_DIVERSIFY(k) is that, they both calculate a weighted coverage of the object categories. It is more practically useful to users. The reason is that users can decide the facet to pay more attention on. Possible facets include aspects, categories, or interpretations of the query according to their own information needs or interests. For example, still on the search of "Harry Potter", users may decide they are interested in books at a ratio of 20 %, while information on movies at a ratio of 80 %. This advantage has also aroused researchers' interest.

MAXUTILITY_DIVERSIFY(k) definition A disadvantage of DIVERSIFY(k) and the QL_DIVERSIFY(k) is that the objective functions in them fail to consider the number of categories covered by the final result set. Therefore, they take into consideration the degree that an object satisfies a certain category, so if a dominant category is not covered adequately, more objects relevant to this dominant category will be selected even though this will pose a threat to other categories.

For this disadvantage, Capannini et al. [9] propose MAXUTILITY_DIVERSIFY(k) to maximize the number of utilities for R while guaranteeing that, diverse specializations of the given query can be covered proportionally. The formal definition is shown as follows.

We use $P(o|q)$ to represent the probability that an object o is selected according to q. Then the MAXUTILITY_DIVERSIFY(k) problem intends to find $R \subset X_q$ with $|R| = k$ to maximize the sum of various utilities for R, i.e., $\widetilde{U}(R|q) = \sum_{o \in R} \sum_{c_i \in C} (1-\theta) \cdot P(o|q) + \theta P(c_i|q) \cdot \widetilde{U}(r|O_{c_i})$. There are three extra constraints: (i) each specialization of the given query is covered in proportion to its probability; (ii) the natural join between O_q and a specialization c_i is equal to a set $\{o \in X_q | U(o|X_{c_i}) > 0\}$; (iii) each specialization $c_i \in C$, the cardinality of the natural join between X_q and a specialization c_i is not less than $\lfloor k \cdot P(c_i|q) \rfloor$.

The advantage of the MAXUTILITY_DIVERSIFY(k) is that it takes into consideration the probability proportion of each specialization of the query. This is useful for users in that it can lead them to know something new, but they may not expect. Still, if we consider the example of search “Harry Potter”, suppose a certain user only takes interests in movies, however, as books are searched and purchased by a lot of other users, this specialization may have a larger probability proportion. Therefore, this user will obtain much information of books, and then, he may start to pay attention to this facet. This phenomenon is really beneficial to the publisher.

xQuAD_DIVERSIFY(k) definition Santos et al. [83] present a framework xQuAD (eXplicit Query Aspect Diversification) to describe query result diversification. Formally, given a user query q, $R \subset X_q$ with $|R| = k$ is to be obtained with each $o_i \in R$ maximizes $M = (1-\lambda)P(o_i|q) + \lambda P(o_i, \bar{R}|q)$ in each iteration. $P(x_i|q)$ represents the likelihood that object o_i is retrieved for a query q, and is related to the relevance of each object. $P(x_i, \bar{R}|q)$ denotes the likelihood that x_i is retrieved according to a query q but not in R, and is related to the diversity between objects in R.

Similarly, HuQuAD_DIVERSIFY(k) [53] is proposed as a hierarchical version of xQuAD_DIVERSIFY(k), with the intent in hierarchical structure.

Obviously, xQuAD_DIVERSIFY(k) is a tradeoff between relevance and diversity. xQuAD_DIVERSIFY (k) and its variants are used in [96] and some other systems. In DivDB [96], users are able to modulate the tradeoff value λ, and present results in low diversification, moderate diversification, as well as high diversification respectively.

Application in the context of documents In Liu et al. [67], the definition of intent-based diversification is used to highlight diverse concepts in the context of documents, such that a notion of cover of a set of sentences is defined in order to describe intent-based diversification in the context of documents. Given a subset $R = \{s_1, s_2, \ldots, s_n\} \subset U$, suppose the sentence $s_i = \{t_1, t_2, \ldots, t_m\}$, where t_j represents a term in this sentence. Then the cover of a sentence set is all the terms appearing in S. That is, Cover$(R) = \bigcup_{s_i \in S} \bigcup_{t_j \in s_i} t_j$. Thus, with regard to a set of documents R with a fixed cardinality, we evaluate its degree of diversity by computing and comparing the number of terms in Cover(K). The more terms Cover(K) has, the more diverse K is. However, this method only takes into account the appearance of each term, but neglects the fact that each term may be covered more than once and different terms may have various contribution weights to the final results.

In order to solve this problem, the times that related terms are covered are taken into account when considering the diversity of the result. Formally, function $f(x)$ and $w(t)$ denote the benefits gained when covering a term x times and the weighted value of term t, respectively. Then the gain of a set R is defined as $\text{gain}(R) = \sum_{x=0}^{|R|} \sum_{t \in \psi_x} w(t) f(x)$, where ψ_x is the set of terms appearing in exactly x sentences in R. $\text{gain}(R)$ is considered as a weighted form of $\text{Cover}(R)$, and therefore, the larger $\text{gain}(R)$ is, the more diverse the document set R is.

2.3 Novelty-based diversification

Novelty-based diversification aims to provide users information which is not contained in previously retrieved objects. Novelty is a notion relevant to diversity. To check whether an object is novel and to improve novelty-based diversity is essential.

Definition based on information nuggets Clarke et al. [16] propose a definition based on information nuggets using the probability ranking principle (PRP) [62,82]. Here, the nugget is a unit to describe how much information a certain object has and this unit is widely used as a result of its utility of measuring information. Given S as the set all the possible information nuggets, the user's information needs can be modeled as a set $I \subseteq S$, and similarly, the information contained in an object o is modeled as a set $\mu_o \subseteq S$.

The objective is to select both diverse and novel objects into R. Obviously, the probability that the kth object is selected is equal to the probability of this document containing an information nugget not found in the previous $(k-1)$ objects. Only in this sense, the kth object is novel and provides some new information not previously mentioned. With a ranked list of $(k-1)$ preceding objects denoted as $(o_1, o_2, \ldots, o_{k-1})$, the probability that a nugget s_i is novel to a certain user query q is $P(s_i \in I | I, o_1, o_2, \ldots, o_{k-1}) = P(s_i \in I) \prod_{j=1}^{k-1} P(s_i \notin \mu_{o_j})$.

We define the number of objects ranked up to $\varnothing_{k-1}$ which have been judged to contain nugget s_i as $\eta_{i,k-1} = \sum_{j=1}^{k-1} J(o_j, i)$, where $J(o, i)$ is a 0–1 variable to describe whether the human assessor judges that o contains s_i. If $J(o, i) = 1$, $P(s_i \in \mu) = \alpha$; otherwise, set $P(s_i \in \mu) = 0$, where $\alpha \in (0, 1]$ is a constant reflecting the probability that the human assessor is erroneous. Assuming that information nuggets are independent and equally likely to be relevant to the user's information needs, then $P(s_i \in I)$ is the same for all s_i, denoted by γ. Thus, the possibility of selection o_k is $P(X_k = 1 | I, o_1, o_2, \ldots, o_{k-1}) = 1 - \prod_{i=1}^{m} (1 - \gamma \alpha J(\varnothing_k, i)(1-\alpha)^{\eta_{i,k-1}})$.

We take a search experience in http://scholar.google.com as an example. We use "database" as the keyword. Suppose, we have obtained several query result sets with a fixed cardinality k. Thus, here we have these result sets qualified in terms of the diversity in hand. Our intention is to calculate each document's novelty-based diversity and pick out the best result set from them. Thus, according to the computation steps, we start to calculate these result sets' values of $P(X_k = 1 | I, o_1, o_2, \ldots, o_k)$. Finally, we choose the result set with the maximum value and consider this as the best in terms of novelty-based diversity of all.

Definition in the context of redundancy measures In Zhang et al. [106], Zhang et al. lay emphasis on the task of identifying novel, redundant objects and then propose novelty and redundancy detection in adaptive filtering. In the research, novelty and redundancy are defined, firstly over a set of objects which are relevant; secondly, according to previously seen objects; thirdly, as opposite endpoints of a scale.

As demonstrated in [22,106], redundancy is measured based on the distance between the new object and previously seen objects. The redundancy of a object o in terms of R is computed as $D(o|R) = \arg\max_{o_i \in R} \text{red}(o|o_i)$, where the redundancy between o and o_i is computed by $\text{red}(o|o_i)$.

Three specialized methods can be utilized to describe $R(d_x|d_y)$, namely set difference, geometric distance and distributional distance.

- *Set difference* this measure represents each object o_i as a set of terms (or nuggets) denoted as $\text{Set}(o_i)$. Then the redundancy is represented as $R(o_i|o_j) = |\text{Set}(o_i) \cap \overline{\text{Set}(o_j)}|$. Note that a term $t \in \text{Set}(o)$ if and only if the term t appears in o more than λ times, where λ is a constant threshold. That is to say $t \in \text{Set}(o)$ if and only if $\text{Count}(t, o) > \lambda$, where $\text{Count}(t, o)$ is the weighted sum of the frequency of t appearing in o, the number of filtered documents containing t, and the number of delivered relevant documents containing t.
- *Geometric distance* some kinds of geometric distance measures are utilized in database research, such as Cosine distance and Manhattan distance [63]. In Zhang et al. [106], Cosine distance, which is a symmetric measure related to the angle between two vectors [55], is used to describe redundancy. In this measure, an object o is represented as a vector, denoted as $o = (\theta_1(o), \theta_1(o), \ldots, \theta_n(o))^{\mathrm{T}}$, where $\theta_i(o)$ is the ith entry of d. Then $R(o_i|o_j) = \cos(o_i, o_j) = \dfrac{o_i^{\mathrm{T}} \cdot o_j}{|o_i| \cdot |o_j|} = \dfrac{\sum_{i=1}^{n} \theta_i(o_i) \cdot \theta_i(o_j)}{|o_i| \cdot |o_j|}$.
- *Distributional similarity* this distributional similarity measure is based on probabilistic language models proposed for identifying relevant documents in information retrieval tasks [59,70,105].
- *Kullback–Leibler Divergence* An object is described by a unigram word distribution τ_d. Then Kullback–Leibler divergence is used to measure the redundancy denoted as $R(o_i|o_j) = -KL(\tau_{o_i}, \tau_{o_j}) = -\sum_t P(t|\tau_{o_i}) \cdot \log\left(\dfrac{P(t|\tau_{o_j})}{P(t|\tau_{o_i})}\right)$. The language model τ_o can be obtained and computed in maximum likelihood estimation (MLE) by $P(t_x|o) = \dfrac{tf(t_x, o)}{\sum_{t_y} tf(t_y, o)}$. As demonstrated in [106], in order to make Kullback–Leibler divergence measure more appropriate, it is necessary to utilize smoothing techniques, such as Bayesian smoothing using Dirichlet Priors and smoothing using shrinkage, to adjust the maximum likelihood estimation.

Difference between intent-based and novelty-based diversity Even though intent-based and novelty-based diversification are both proposed to increase the overall diversity of the final result set, they lay emphasis on different facets. Intent-based diversity meets the requirement of resolving ambiguity to obtain a comprehensive understanding of the user's needs, whereas the goal of novelty is to avoid redundancy to make sure that the user can access various information resources [16]. In Clarke et al. [16], a framework is proposed to give a precise distinction between intent-based diversity and novelty, and thus, it can improve the quality of the final query results. In this framework, documents are concretely linked to relevance through informational nuggets, which can represent both the properties of documents and an information need's components. Then this framework is developed into a specific evaluation measure on the basis of cumulative gain.

Novelty-based Diversity and Redundancy In some cases, increasing novelty will lead to the redundancy, some definitions are proposed to promote novelty without resulting in redundancy.

Zhang et al. [106] demonstrate the importance of extending an adaptive information filtering system to make decisions about relevant documents' novelty and redundancy. Then they propose the idea that relevance and redundancy (related to novelty) should be modeled separately. An information filtering system is considered well qualified if, for each user profile, it is able to identify three categories: (i) not relevant documents, (ii) relevant documents containing no new information, i.e., relevant but not novel, (iii) relevant documents containing some new information, i.e., both relevant and novel.

Five redundancy measures are proposed in [106], namely, set distance, Cosine distance, Shrinkage language model, Dirichlet Prior language model, and Mixture language model. Then an evaluation methodology is established, and by using this methodology, these five redundancy measures are thoroughly analyzed.

Intent-based, novelty-based and content-based diversification can be used together. We can first use content-based method to balance the tradeoff between relevance and diversity. Then, we can make use of intent-based diversity to meet the requirement of resolving ambiguity to obtain a comprehensive understanding of the users needs. Finally, novelty-based diversification is used to increase the overall diversity of the final result set. In this way, we can make sure that the user can access various information resources.

3 Content-based diversification algorithms

The content-based diversification problem can be viewed as a bi-criteria optimization problem with the objective to balance the tradeoff between relevance and diversity. A lot of algorithms have been proposed to solve this problem, and generally, they can be categorized into two groups: the first group is based on interchange operations, i.e., swap operations, so we call them interchange algorithms. The second group is greedy algorithms. We introduce interchange and greedy algorithms in Sects. 3.1 and 3.2, respectively. After that, we discuss the categorizations of the approaches in Sect. 3.3.

3.1 Interchange algorithms

Interchange algorithms aim to increase result quality by swapping some object in result set with a better one. The combination of $\delta_{\text{sim}}(q, o_i)$ and $\delta_{\text{div}}(o_i, o_j)$ are used as the objective function F.

Swap method Swap method [104] generates R from U to maximize F, such that the objects in R are not only relevant to the query, but are as diverse as possible. This algorithm keeps on swapping the object who contributes least to diversity, with the next most relevant object in the remaining objects until all objects in U are considered. Initially, R is set to be top-k relevant elements in U and the objects in $U-R$ are stored in a descending order according to $\delta_{\text{sim}}(\cdot, \cdot)$ values. In each of the following iterations, all objects in R are checked and the one $o_i \in R$ with the minimum diversity is swapped with $o_j \in U-R$ with the largest $\delta_{\text{sim}}(q, o_i)$, if such swapping could increase F.

Swap method is an intuitive and simple algorithm using interchange operations. However, $\delta_{\text{div}}(\cdot, \cdot)$ is not considered in the choice of the objects in the set U. As a result, it cannot guarantee that the maximal F will be achieved due to the influence of the $\delta_{\text{div}}(\cdot, \cdot)$.

BSwap method BSwap method [104] shares the same framework as Swap method. The difference is that in each iteration, it is ensured that the sum of $\delta_{\text{div}}(\cdot, \cdot)$ is increased, such that the objects in the result set are as different from each other as possible, i.e., the result set is as diverse as possible. In this method, interchange operations are carried out as long as the value of the sum of $\delta_{\text{div}}(\cdot, \cdot)$ values is increased.

The weakness of this BSwap method is that the retrieved objects may not be relevant enough to the user query request since only the diversity is considered during swapping and the iterations with all objects in R lead to high computation complexity. Additionally, a proper threshold used in the algorithm is difficult to choose and an improper threshold will cause low result quality.

3.2 Greedy algorithms

Greedy algorithms construct the result set by selecting the "optimal" object from the universe incrementally according to some criterion. The objective is still to maximize F.

Maximal marginal relevance (MMR) method Carbonell and Goldstein [10] propose "marginal relevance" to describe both relevance and diversity. Thus, an object with maximal marginal relevance (MMR for short) is both relevant to the user query and different from the objects in current R. MMR is computed as $\text{MMR} = \arg\max_{o_i \in X \setminus R}[\lambda(\delta_{\text{sim}}(o_i, q) - (1 - \lambda)\max_{o_j \in R}\delta_{\text{div}}(o_i, o_j))]$. Maximal marginal relevance method constructs the result set by selecting a new object according to MMR iteratively till no element could be added to R.

The disadvantages of this approach is that the quality of R is sensitive to the first selected object. If the first object is not chosen properly, the algorithm's performance will suffer.

Motley method Jain et al. [54] present a Motley method using greedy strategy of adding the object with diversity value to each object in R greater than a predefined threshold γ to R. This process continues until $|R| = k$ or the candidate set gets empty.

Motley method uses a threshold value γ to control whether an object will be added to R. However, a proper γ is difficult to choose. Motley method also shares similarity with maximal marginal relevance method in that the first selected object has great influence on R, and it is also of vital importance to choose the first object properly.

Max-sum dispersion [32] proposes a 2-approximation greedy algorithm to solve max-sum dispersion problem. In this approach, the distance metric between two objects o_i and o_j is defined as $\text{dis}(o_i, o_j) = (1-\lambda)(\delta_{\text{sim}}(q, o_i) + \delta_{\text{sim}}(q, o_j)) + 2\lambda\delta_{\text{div}}(o_i, o_j)$. In each iteration, the pair of objects o_i and o_j with maximal $\text{dis}(o_i, o_j)$ is selected into R.

This method skilfully balances relevance and diversity between with the distance metric, and the metric is easy to compute. However, as this method is used with regard to a pair of objects, it requires even objects in the result set. Therefore, its performance will suffer if the desired size of the final result set is odd. In that case, an extra object has to be selected randomly from the candidate set, which will affect the result set and finally, the users' satisfaction.

Clustering-based method A clustering-based method is proposed and applied in [93]. Such method is divided into two phases. In the first phase, the k-medoid algorithm clusters X into k clusters according to $\delta_{\text{div}}(o_i, o_j)$. Then in the second phase, this method aims to pick out an object from each cluster to construct R. The disadvantage of this method is that the tradeoff between relevance and diversity could only be considered in the second phase.

Greedy marginal contribution (GMC) method As an improvements of the MMR method. Vieira et al. [95] propose greedy marginal contribution method (known as GMC). R_{p-1} denotes the partial result set with size $p-1$, and $1 \le p \le |R|$. For an object o_i, its maximum marginal contribution $\text{mmc}(o_i)$ is defined using three components. The first is $\delta_{\text{sim}}(q, o_i)$, the second is the diversity between o_i and other objects in R_{p-1}, and the third component is the diversity between o_i and objects in $X-R$. $\text{mmc}(o_i)$ is a summarization of above three components. The GMC algorithm constructs the result set R by selecting the object with the highest mmc value incrementally and greedily.

The choice of the first object has great influence on R and the algorithm's performance. GMC algorithm takes into account both the relevance factor and the diversity factor, but MMR algorithm only considers the relevance factor. In this sense, the GMC algorithm outperforms the MMR algorithm.

Greedy randomized with neighborhood expansion (GNE) method In Vieira et al. [95], greedy randomized with neighborhood expansion method (known as GNE) is introduced as a combination of greedy and swap approaches. GNE algorithm uses the greedy randomized adaptive search procedure (GRASP) technique during the process of diversifying the results. This is the first randomized solution for the diversification problem. GNE algorithm consists of two phases, GNE construction phase and GNE local search phase. The first one selects R according to a greedy randomized ranking function with mmc embedded. The second one improves R by swapping between some element in R and the most diverse element in $X-R$ to another element in R. Its difference to GMC is that it selects a random object among the top ranked objects according to their mmc values. Such randomized operations accelerate the process, but may miss the optimal solution.

DIVGEN Algorithm Angel et al. [3] propose a threshold algorithm to solve the diversity-aware search problem. This algorithm utilizes novel data access primitives and improves the performance significantly. It offers performance benefits as it maintains bounds on the candidate objects (for example, documents) and bounds on the usefulness of these objects. Thus, the algorithm is able to reduce the number of objects examined during the process.

An Algorithm with monotone submodular set functions Borodin et al. [5] conduct a research on a particular version of diversification, max-sum diversification (based on objective functions). The max-sum diversification is defined to pick out an object subset from all the objects in order to maximize the sum of this subset's relevance to the query and a diversity measure of this subset. A 2-approximation greedy algorithm for max-sum diversification problem is proposed with monotone submodular set functions, which are valuation functions more generally than others. Then this problem is further extended to matroids and a 2-approximation local search method is proposed.

Dos-Naive and Dos-overlap Methods Khan et al. [58] propose the DoS (diversification of multiple search results) scheme that addresses the problem of efficiently diversifying the results of multiple queries. Toward this goal, DoS leverages the natural overlap in search results in conjunction with the concurrent diversification of those overlapping results. This enables DoS to provide the same quality of diversification as that of the sequential methods, while significantly reducing the processing costs.

3.3 Analysis about the categorization of various approaches

The proposed algorithms are demonstrated in the "Interchange/Greedy" column in Table 1. Although this classification method is natural, we cannot ignore two other useful categorizations as follows.

The second categorization is based on whether an algorithm uses a threshold parameter when constructing the final result set. When a threshold parameter is used, it cannot be avoided that the algorithm's performance is affected by some extent since choosing a proper threshold is difficult. In this sense, the proposed algorithms can be divided into two categories and this is shown in "Use a Threshold Parameter (Yes or No)" column in Table 1.

Table 1 The categorization of content-based diversification methods

Algorithm	Interchange/ greedy	Use a threshold parameter	Use a balance parameter
Swap method [104]	Interchange	No	No
BSwap method [104]	Interchange	Yes	No
Maximal marginal relevance (MMR) [104] method	Greedy	No	Yes
Motley method [54]	Greedy	Yes	No
Max-sum dispersion [32]	Greedy	No	Yes
Clustering-based method [93]	Greedy	No	Yes
Greedy marginal contribution method [95]	Greedy	No	Yes
Greedy randomized with neighborhood expansion method [95]	Greedy	No	Yes
DIVGEN algorithm [3]	Greedy	Yes	No
Algorithm with monotone submodular set functions [5]	Greedy	No	Yes

Then the third categorization focuses on whether an algorithm uses a parameter in the objective function to efficiently balance relevance and diversity. It is obvious that using a threshold will help balance the degree of relevance and diversity and can also decide whether the relevance or diversity is more important in obtaining the final results. According to this criterion, we categorize existing methods into two groups. One uses a balance parameter, while the other does not. These two groups are illustrated in the "Use a Balance Parameter" column in Table 1.

After respectively presenting three ways to categorize the content-based diversification methods, we provide a global picture of all the dimensions which can be used to analyze and assess these algorithms. The detailed categorization dimensions are shown in Table 1.

4 Intent-based diversification algorithms

Besides the basic settings, the settings of intent-based diversification algorithm also include the taxonomy and the probability that a category is relevant to the query. Intent-based diversification algorithms have three kinds. The first kind aims to carry out query–query reformulations to better improve the diversity of user intents. The second kind is based on mining specializations from query logs and then re-ranks original results to diversify the results. The third kind is to compute a final ranking on the basis of both relevance and dissimilarity. Each of these three approaches has its own features. When trying to accomplish the task of intent-based diversification, we should choose the proper approach according to its pros and cons.

4.1 Three algorithms with query–query reformulation

As stated above, different users may expect different information even given the same query, especially when the query request is ambiguous. Therefore, diversifying the query results to satisfy users' various intents is of vital importance.

To avoid the homogeneity in query intents, Radlinski et al. [79] propose a framework using query–query reformulations to understand the diversity of user intents. Then based on this framework, three algorithms are presented in [79]. This problem is formally defined as follows. Given a query q, generate a set $M(q)$ of k queries both relevant to q and different from q, and thus, these queries may reflect various user intents. In order to accomplish query–query reformulations, a large sample of the query logs is obtained and thoroughly analyzed from a popular web search engine. Then three algorithms with query–query reformulations are proposed as follows.

The first algorithm, named most frequent method, sets $M(q)$ as the reformulations of query q. Note that the queries in $M(q)$ have the highest number of times following q. The second algorithm, called maximum result variety method, selects the queries which are not only frequent reformulations but also different from other reformulations already selected to construct $M(q)$ greedily. The third algorithm, most satisfied method, sets $M(q)$ using the queries which tend not to be further reformulated and have a certain frequency (both controlled by predefined thresholds).

These three methods have their own features. The first one emphasizes the frequency that a query reformulation follows q. This method fails to take into consideration the diversity between reformulations. In this sense, it fails to balance the relevance and diversity properly. The second method revises the problem in the first method, by making sure that the reformulations selected are different from each other, but it should be guaranteed that reformulations selected cannot be too irrelevant to the original query q_i. Otherwise, users will also be unsatisfied when they get too much information not relevant enough to their requests. The third algorithm further revises this exposed disadvantage in the second method, by setting a predefined threshold to avoid this irrelevance problem. However, it is hard to choose a proper threshold, which has a deep influence on the final result set.

4.2 OptSelect based on query logs

Apart from these three algorithms mentioned above, Capannini et al. [9] present OptSelect algorithm to promote intent-based diversification. It re-ranks the original results through mining specializations from query logs. Once a user query q is submitted, the algorithm performs the following three steps: (i) check whether q is ambiguous, that is to say, if q can be divided and analyzed in different facets, and if so (ii) mine specializations from query logs and retrieve documents for all of them in order to collect information about the different specializations; finally, (iii) with the aim to maximize the possibility to satisfy the users' needs, a final result set is built using the relative frequencies of the specializations mined.

This algorithm requires the developers to mine specializations from query logs. This mining job is not easy and may be both time-consuming and resource-consuming and result in high computation overhead.

4.3 Topic diversification algorithm

Ziegler et al. propose a new topic diversification algorithm to consider the users' full range of intents when they are offered. In this way, relevance and diversity are both taken into account. First, we have an original rank denoted as P_{org}. Then, the dissimilarity rank P_{dis} is obtained by sorting all the objects according to a certain rule. Finally, given a diversification factor θ_{F}, which represents the impact of P_{dis} exerted on the final ranking, we merge P_{org} and P_{dis} according to θ_{F} and then obtain the final ranking P_{final} of the objects.

Obviously, this algorithm manages to consider both relevance and diversity. However, we notice that the diversification factor θ_{F} has a great influence on the final ranking and performance. Therefore, we have to choose this diversification factor properly. It is really not an easy task.

5 Diversification methods on specialized applications

The methods in Sects. 3 and 4 could be applied to all kinds of search tasks with corresponding settings, even though some of them are proposed for some specialized search. Different from them, some methods are specially proposed for some kind of query. In this section, we describe and analyze such approaches, which are classified according to the query type and data type to handle.

5.1 Diversification in query over structured data

5.1.1 Diversification in top-k queries

Top-k query processing has been studied thoroughly and deeply in the past. This problem aims to retrieve several results to form a list maximizing the overall score with regard to a user query request. All these results are independent. However, in practical study, they can be similar to each other. Thus, the user's satisfaction can benefit from introducing diversification in the top-k problem.

Qin et al. [78] formalize the diversified top-k search problem. Then two categories of algorithms, the incremental top-k framework and the bounding top-k framework which are generally used in the top-k problem with early stop property, are studied. Then these two frameworks are both extended to solve the diversified top-k search problem by adding three application-independent problems whose solutions are represented as functions, namely, a sufficient stop condition denoted as sufficient(), a necessary stop condition denoted as necessary(), and a diversity search function denoted as div-search-current(). Through introducing these three functions, we can improve the efficiency to a great extent. Therefore, the diversified top-k search problem can be better solved. Then it is shown that div-search-current() is an NP-hard problem and is hard to be approximated, so three new algorithms, div-astar, div-dp, div-cut, are proposed to find the optimal solution for div-search-current(). Finally, extensive studies are conducted to test the performance with two real datasets and experimental results prove these newly introduced algorithms efficient.

There is also a special kind of top-k diversity queries which intend to retrieve the best *k* objects which are not only relevant to the query, but also well distributed in a designated region. The objects for querying are embedded in a low-dimensional vector space. This special top-k diversification is demonstrated in [11]. This paper is the first one to point out the fact that this kind of diversification does not need accessing and scanning all relevant objects to find the best *k* objects. Also, it is the first work which introduces bounded diversification with sorted access methods. Then Catallo et al. propose an algorithm called space partitioning and probing (SPP), which progressively explores the vector space, and tracks the already seen objects with their relevance and position. In this way, the goal to return a qualified result set in terms of both relevance and diversity can be achieved.

Apart from these works on diversification of top-k problem mentioned above, there are also some variants of top-k problem with their corresponding solutions. For instance, Nanongkai et al. [72] put forward the k-regret (short for k-representative regret minimization query) and utilize it to support multi-criteria decision making. Concretely, Nanongkai et al. propose a new definition named the *maximum regret ratio* which describes how disappointed the users would be if he had seen the *k* representative results rather than the whole database. Then for any number *k* and any class of utility functions, the output of the k-regret query is *k* tuples from database which can minimize the maximum regret ratio. Furthermore, they have proved that the maximum regret ratio can be bounded and the bound is not dependent on the database size.

5.1.2 Diversification in multi-dimensional search queries

Although most existing works on diversification is based on a specific attribute, it is also important to increase the diversity of returned results for queries with constraints on multiple dimensions over relational database.

Dou et al. first carry out some subtopic mining from four sources of data, namely, anchor texts, query logs, search result clusters and web sites [20]. Then Dou et al. propose a general framework to diversify the final results taking into consideration various dimension of subtopics. This framework can well balance the relevance of documents to the query and the diversity (can be regarded as intent-based and novelty-based diversity) between documents. Finally, two approaches such as a topic richness model and a topic novelty model are implemented to diversify the results concretely.

Liu et al. give a definition of the differentiability (which can be seen as diversity) of structured query results as well as differentiation feature set (DFS) for each result, and then quantify the degree of difference [66]. Then three desiderata: differentiability, validity and small size are identified to obtain good DFSs. Next, a theoretical conclusion is drawn that the problem of identifying a limited set of features to diversity the result set to the most extent is NP-hard. As a result, in order to get practical solutions, two local optimality criteria, namely single-swap optimality and multi-swap optimality, are proposed and then, several efficient algorithms are designed to meet these criteria.

5.1.3 Diversification in keyword search over relational databases

Stefanidis et al. [87] focus on this facet of diversification study and make the following contributions. First, on the basis of database keyword search, they manage to provide a formal model to integrate user preferences into the final ranking. Second, in this model, Stefanidis et al. combine the relevance, each individual result's degree of preference, user interest coverage (i.e., intent-based diversity) and content diversity (i.e., content-based diversity) together to comprehensively evaluate the quality of returned results. Third, based on this combined criterion, they provide efficient algorithms in order to compute top-k representative results.

Then in [17], Demidova et al. propose a novel approach named DivQ to diversify the returned results over structured databases in the area of keyword search. To begin with, they introduce a probabilistic query disambiguation model which is used to rank the possible interpretations of a keyword query on structured databases. Next, a diversification scheme is proposed which re-ranks the query interpretations considering redundancy of query results. In this scheme, additionally, a greedy algorithm to select diverse query interpretations is put forward. Finally, in order to evaluate the quality of diversification in structured database keyword search, two metrics, α-nDCG-W (an adaptation of α-nDCG) and WS-recall (an weighted adaptation of S-recall) are, respectively, proposed. Through experimental evaluation, it is proved that the search results obtained by this proposed method can better characterize possible answers available in the database than initial relevance ranking.

5.1.4 Diversification regarded as a spatial concept

In some existing works, diversification is often regarded as a spatial concept which is related to distance or proximity. In Gonzalez [33], the problem of clustering a set of points so as to minimize the maximum intercluster distance is studied and then an algorithm which can guarantee solutions with an objective function value no larger than two times the optimal solution value is given.

In Multi-dimensional space In multi-dimensional or high-dimensional space, it is also studied to make the final returned result set more diverse when processing queries, in the area of fast nearest neighbor search [4], in geographical web search engines [14], and in geographical information retrieval with scattered ranking methods [92].

In Spatial search In the area of spatial search, there are also some existing and meaningful works. For instance, Ni et al. focus on point-wise region queries [74], and then they present an exact method as well as an approximate method to deal with this problem. Cao et al. pay attention to spatial keyword querying [8] and give a thorough survey on various keyword querying functionality with the ideas underlying their definitions. Then Tang et al. lay their research emphasis on the evaluation as well as user preference on spatial diversity [88], and they manage to show the potentials of spatial diversity through a user preference study using Amazon Mechanical Turk.

When diversification is viewed as a spatial concept related to proximity, proximity rank join problem [69] is proposed. In this problem, the score function reflects the diversification. Such problem pays attentions on the multiple relations. The ProxRJ algorithm for this problem is an iterative algorithm. In each iteration, the pulling strategy is applied to decide the next relation R to be accessed. Then the next unseen tuple in R is joined with other relations and top-k results are added to the output set. At the end of the iteration, the upper bound on the aggregate score of the unseen combinations is updated. The algorithm halts when the smallest score of results lower than the bound or the size of the results reaches the given k. This algorithm is based on a tight upper bound which can guarantee instance-optimality.

5.2 Diversification on graphs search and mining

In the area of graph research, there are a lot of problems which need to compute all or top-k subgraphs. As these subgraphs are likely to overlap with each other, the diversification methods are of vital importance to such graph problems. In this part, we will focus the diversification methods on graphs.

On graph pattern matching In Fan et al. [29], Fan et al.lay their emphasis on the graph pattern matching problem. By supporting a designated output node, they revise graph pattern matching problem and then study two function classes to rank the matches. Additionally, two algorithms are developed for computing top-k matches and then diversified top-k matching, which is a bi-criteria optimization problem considering both relevance and distance, is well studied.

On redundancy-aware search Wang et al. [98] study how to obtain redundancy-maximal cliques. They try to provide a summary of the set of maximal cliques and make sure the summary is concise and complete. To be specific, they introduce the notion of τ-visible MCE (maximal clique enumeration) to both reduce the redundancy and capture the result's major information. Then Wang et al. also propose efficient algorithms to compute a τ-visible summary as well as introduce top-k clique computation based on the summary. In this way, the result usability can be further enhanced.

In Xin et al. [101], focus on how to extract redundancy-aware top-k patterns. After examining two problem formulations: MAS (maximal average significance) and MMS (Maximal Marginal Significance), MMS is found more reasonable for the formulation of this problem.

They next present an improved algorithm, whose performance is bounded by $O(\log k)$ for MMS. Then through the evaluation of two case studies as well as experiments, the proposed method's performance is examined.

On graph feature selection in graph classification Zhu et al. [108] pay their attention to the problem of graph classification and then provide a diversified discriminative feature selection approach. After a deep study on the most widely used selection approach, Zhu et al. design their feature selection approach by introducing a new diversified score which can reduce the overlap between selected features. Then through experimental evaluation, it is demonstrated that the introduced diversified discriminative score is able to make positive graphs and negative graphs separable and thus contribute to a higher classification accuracy.

On node ranking in large graphs Li et al. [65] make some contributions to the scalable diversified ranking on large graphs. To begin with, they propose a novel diversified ranking measure, which takes into consideration both relevance and diversity. Then this proposed measure's submodularity is proved and an efficient greedy algorithm with good performance is designed to achieve near-optimal diversified ranking. Finally, Li et al. present a generalized diversified ranking measure and develop an efficient randomized greedy algorithm to accurately obtain its maximal value.

Similarly, Tong et al. [89] research the diversified ranking on large graphs from an optimization viewpoint. First, a goodness measure which considers both relevance and diversity is proposed for a top-k ranking list. Then a scalable algorithm named DRAGON which is linear to the graph, is proposed and then it is proved that DRAGON can generate a provably near-optimal solution.

On top-k structural search As the top-k problem also exists in the area of social networking and social data analysis, diversification begins to play an essential role in these areas. Fan et al. [29] study the diversified top-k graph pattern matching in order to better analyze social data. First, the traditional notion of graph pattern matching is revised by designating an output node to prepare for the following research. Then two classes of functions, namely, $\delta_{\text{rel}}(\cdot,\cdot)$ called relevance functions which measure the relevance of a certain match, and $\delta_{\text{dis}}(\cdot,\cdot)$ called distance functions which measure the distance, i.e., dissimilarity between two matches, are proposed. With a balance parameter λ, a bi-criteria diversity function denoted as $f()$, aiming to take both relevance and diversity into consideration, is defined to maximize social impact and cover social elements as diverse as possible. Then based on $f()$, the diversified top-k graph pattern matching problem is turned into a bi-criteria optimization problem. Both an approximation algorithm and a heuristic algorithm with the early termination property are given to solve this problem.

Huang et al. [52] present their research on top-k diversity search in large networks. In order to solve this problem, first, a novel top-k search framework is developed, with a Union-Find-Isolate data structure to keep track of obtained structural information already and an effective bound used for pruning. Then an A*-search-based algorithm is proposed, according to a heuristic search order devised to traverse the components in a vertex's neighborhood, to compute the structural diversity of a vertex. Additionally, some techniques are designed and utilized to handle updates in dynamic networks and keep the top-k results maintained.

[102] studies the diversified top-k clique search problem. The goal is to find top-k maximal cliques that can cover the most number of nodes in the graph. For memory issues, this approach only maintains k candidates during the maximal clique enumeration to achieve memory bound. A light weight online index is developed to reduce the time complexity

furthermore. Additionally, global pruning, local pruning, and initial candidate computation are designed to improve the efficiency. Global pruning determines a global search order of nodes to terminate search on useless nodes. Local pruning is used to avoid expanding unpromising partial maximal cliques. Initial candidate computation initializes candidates using a greedy strategy.

5.3 Diversification in search over unstructured data

Diversification is really important and widely applied in the areas of web search and information retrieval. For instance, in [79], Radlinski et al. propose several methods to diversify results according to a query using past query reformulations. Given a user query, the objective is to generate a set of related queries with a fixed cardinality. Radlinski and Dumais [79] provides three ways to achieve this goal, namely, the most frequent (MF) method, the maximum result variety (MRV) method, and the most satisfied (MS) method. Specifically, the MF method selects queries which follow the given user query most often, so it fails to consider the diversity of the returned set. Then the MRV method overcomes this by selecting reformulations both frequent and dissimilar to the reformulations which have been returned. However, the MRV method fails to guarantee that the returned reformulations cannot be too irrelevant to the query. To revise this, the MS method sets a minimum frequency to make sure that the returned reformulations are relevant. Then after presenting three ways to increase the diversity of the top results, the effectiveness of these methods is thoroughly evaluated.

Moreover, some other researchers also pay much attention to the diversification in this area. Capannini et al. [9] aim at realizing efficient diversification, which is specific to web search results. To begin with, a methodology is defined about when and how to diversify the query results. Then, a measure is defined to describe a result's value for a diversified result list. Additionally, OptSelect, an original algorithm is presented in order to accomplish the diversification task both effectively and efficiently. Finally, the algorithm's performance is tested on the standard TREC Web diversification track testbed against some state-of-the-art diversification methods.

Similarly, in [16], Clarke et al. consider both intent-based diversification and novelty-based diversification for information retrieval systems. A framework is presented for evaluation that rewards novelty and diversity systematically. Then the resulting framework makes it possible to distinguish novelty aiming to reduce redundancy and diversity aiming to resolve ambiguity. Therefore, on the premise that the accuracy of retrieved documents is guaranteed, a balanced gain of both novelty and diversity is obtained.

Also, Rafiei et al. [80] also focus on the study of making the web search results more diverse. Concretely, they manage to model this diversification problem into an optimization task in which the expectation should be maximized. Then, they try to estimate the model parameters such as correlations between pages, relevance expectation, variance and target expectation. Then finally, an equilibrium between these parameters is reached. This is the first work to relate result quality as well as diversity to clicks' expected payoff and risk in the area of web search and then include these quantities in a model to achieve optimization.

5.4 Diversification in recommendation systems

Recommendation systems are utilized to offer recommendation lists of products to users that they may appreciate. This recommending procedure is based on users' past preferences, shopping history, and some demographic information [110]. In order to avoid the homogeneity of the recommendation lists, several diversification methods are proposed. In this section, related works to improve the diversity of recommendation systems are discussed.

Ziegler et al. [110] propose a novel method, named as topic diversification, in order to improve the intent-based diversification in personalized recommendation lists and satisfy the user's needs. This method takes into account both accuracy of recommendations, i.e., relevance, and diversity, based on the assumption that users will not be satisfied if only accuracy is guaranteed. Then topic diversification's implication on item-based [35,81] and user-based [19,85] collaborative filtering are both analyzed thoroughly.

Some specific recommendation systems with the consideration of diversity are developed. DiSCern [12] is a citation recommendation system for a given scientific query topic. DiSCern finds relevant and diversified citations in response to a search query in form of keywords. In DisCern, the citations are modeled as graphs. Communication discovery algorithms are developed to ensure that the results contain semantically correlated articles. Vertex reinforced random walk (VRRW) is adopted to balance prestige and diversity of the results. In the music recommendation system [86], diversity is considered to enhance music recommendation.

[75] proposes clustering-based diverse friend recommendation approach. Density-based fault tolerant subspace clustering approaches are adopted. For efficiency issues, significance threshold is used to prune unpromising subspaces. The diversification goal is achieved by refining like-minded users with a weighted ranking approach.

5.5 Diversification in time series search

Eravci et al. [28] put forward a diversity-based relevance feedback approach according to time series search. It is shown that both relevance feedback and diversity can be beneficial in enhancing the established representation in time series applications. Firstly, it is known that relevance feedback increases the retrieval accuracy, and then, it is found that diversity can contribute to the performance further in many of the cases in the first iteration of the relevance feedback. Eravci et al. try to consider diversity in the area of top-k retrieval and then explore two methods, namely, maximum marginal relevance and cluster-based diversity, to diversify the results. Given an object, maximum marginal relevance takes into consideration both the distance between it and the query, and the distance between it and other existing objects. Well, with the nearest neighbor method, cluster-based diversity manages to retrieve Top-αk elements, then k clusters are obtained from the αk elements. Finally, choose the data points which are nearest to the cluster centers or the predefined representatives are retrieved as the final results. It should be noted that the parameter α describes the diversity desired. If evaluated by precision, cluster-based diversity with certain parameters in [28] is verified to be the best.

Non-pure and non-separable data cases are generally the challenging cases in retrieval systems; however, the content-based diversification methods demonstrated in [28] performs better in these cases where data are not pure or not separable. In this sense, diversification can contribute a lot to the area of time series search.

5.6 Diversification in querying streaming data

Since most existing methods to increase the diversity of the results need random access to input, they are not applicable when the input data is continuous. Thus, it is of vital significance to put forward some diversification approach specific for streaming data.

Minack et al. [71] propose an incremental diversification framework in which the input data is treated as a stream. Additionally, this approach exhibits a linear computation and its memory complexity is constant with respect to the input size. In practice, this method can be used to increase the diversity of large sets as well as continuous data streams with guarantee in applicability and efficiency.

A two-step diversification algorithm in publish/subscribe systems is proposed in [21]. In the first step, given k, l and T, we compute the k most diverse objects in disjoint periods, i.e., different T_i. Then these k objects can be returned to subscribers at the end of a certain period. In the second step, the k most diverse objects are calculated over windows of length l with the help of sliding-window delivery. In this way, we can learn that an object could be forwarded to subscribers from publishers if and only if this object is one of the k most diverse objects in every sliding window it belongs to. Finally, in the third step, a certain new object is delivered by history-based filtering delivery if and only if it is different enough from the k most diverse objects which are recently selected.

Drosou et al. [23,26] study the dynamic diversification of continuous data. In this paper, Drosou et al. try to focus on the problem of selecting k-most diverse objects from the result set in a dynamic setting. First, they define the continuous k-diversify problem with constraints that can make the diversified results continuous. Then an approach based on cover trees is proposed, and this approach is able to support dynamic insertion and deletion of objects. Next, as this kind of diversification problem is generally NP-complete, the authors provide theoretical bounds to evaluate their approach compared with the optimal solution.

Top-k diversity subscription query processing is studied in [13]. Such query considers three aspects of results: (1) relevance; (2) freshness and (3) diversity. Such query is processed on the text stream. For efficiently processing, group and individual filtering techniques are proposed. The former formalizes the problem as a minimal covering set maximization problem and users greedy algorithm to solve such problem. The latter adopts query result index to compute the score and determine whether a document is the result of a query efficiently.

6 Query diversification systems

The final goal of query result diversification is to improve users' satisfaction through certain algorithms, frameworks or models. Thus, it is also essential to improve user experience when users are interacting with web search engines and so on. In this area, some novel tools and systems have been developed to achieve this goal.

For instance, Drosou et al. [25] put forward a tool called POIKILO in order to provide assistance for users in locating and evaluating diverse results. Then, models and algorithms are implemented to compute and compare diverse results. As for users, they can not only access tuning various parameters of diversification process, but also combine diversity with relevance and observe the change over time in the final results when dealing with streaming data. It provides the users with the ability of offering the option to tune the degree of diversification by zooming-in or zooming-out of the presented subset.

In [96] a system named DivDB used for diversifying query results is proposed. As when query parameters change, the performance and the result of each method vary correspondingly, whereas generally users are not aware of which method is the most satisfactory. As for such situations, some heuristic-based rules are developed to help DivDB choose the most appropriate diversification algorithm. DivDB is the first system allowing users to make comparisons among various diversifying algorithms and when tuning the query parameters, it offers a friendly interface for users to inspect the result change. This system can be beneficial to both advanced users and novice users, in that it provides a flexible platform for diversifying query results. A SQL-based extension is provided to formulate queries with diversification for advanced users, as they may be interested in the performance of various diversifying algorithms. However, DivDB also thoughtfully allows the novice users to provide a “hint” to the optimizer on speed versus quality of query result, as the final query results may be more attractive to them than the various algorithms’ parameters which can be tuned.

Specialized in the area of keyword search in databases, a system BROAD is presented in [107]. This system makes it possible that users can perform diverse and hierarchical browsing on keyword search results. To realize this concretely, BROAD first partitions the answer trees in the keyword search results through picking out *k* elements which are representative from the trees. Then the answer trees need to be separated into *k* groups based on their similarity to the representatives selected. In the next step, the partitioning operation is conducted for each group. Finally, in each group, by constructing the summarized result for the answer trees as above, BROAD provides a method for users to locate their desired results quickly.

7 Discussions

In this section, we propose a summary of the approaches and introduce other approaches that improve efficiency and effectiveness.

7.1 Summary of approaches

We compare the definitions of result set diversity, query processing methods with result set diversity as a goal and applications of diversification techniques in Table 2, where “generalized” means that the method is suitable for all cases. For an application, the diversification method could be selected according to the guide in this table.

The approaches could be selected according to the setting. Content-based diversifications are more suitable for the case that the similarity or diversity between objects is easy to compute. Intent-based diversification methods are workable when category or topic information is available. If a set of results exist and novel results are to be updated, novelty-based diversification methods are applicable.

Besides the setting, other factors are to be considered. The first factor is the data model, i.e., structured data, semi-structured data such as graph or unstructured data such as text. Some approaches are suitable for general case, as identified in Table 2. While some approaches specially for some kind of data models are summarized in Sect. 5. The second factor is the ranking requirement, i.e., whether the results should be ranked. If ranking is required, the approaches with special ranking function should be considered and an extra ranking function should be defined if approaches without ranking functions are selected. The last but not least, the efficiency and scalability should be considered since these issues should be considered in database and information retrieval systems. Table 2 gives suggestions. Besides, the excepted running time could be estimated according to the time complexity reported and the experimental results on small-size data set.

As shown in Table 2, we can intuitively know the efficiency and application of different query processing methods. In content-based diversification, BSwap method [104], maximal marginal relevance (MMR) Method [10], motley method [54] and clustering-based method [93] are of high efficiency. Efficiencies of swap method [104] and greedy marginal contribution (GMC) method [95] are medium while max-sum dispersion(MSD) [32] and greedy randomized with neighborhood expansion (GNE) method [95] are of low efficiency. Swap method [104], BSwap method [104], maximal marginal relevance (MMR) method [10], Motley Method [54], greedy marginal contribution (GMC) Method [95], Greedy randomized with neighborhood expansion (GNE) Method [95], DIVGEN Algorithm [3] and

Table 2 Diversification zoo

Definition	Query processing method	Efficiency	Application
Content-based	Swap method [104]	Medium	Generalized
	BSwap method [104]	High	Generalized
	Maximal marginal relevance (MMR) method [10]	High	Generalized
	Motley method [54]	High	Generalized
	Max-sum dispersion (MSD) [32]	Low	Graph
	Clustering-based method [93]	High	Graph
	Greedy marginal contribution (GMC) method [95]	Medium	Generalized
	Greedy randomized with neighborhood expansion (GNE) method [95]	Low	Generalized
	DIVGEN algorithm [3]	–	Top-k and Generalized (diversity-aware search)
	Algorithm with monotone submodular set functions [5]	–	Generalized
Intent-based	Most frequent method [79]	High	Web search
	Maximum result variety method [79]	Medium	Web search
	Most satisfied method [79]	Low	Web search
	OptSelect [9]	High-to-medium	Generalized (query logs required)
	Topic diversification algorithm [110]	Medium	Generalized (ranking required)
Novelty-based	No implementation algorithm	–	–
–	Div-astar [78]	Low	Top-k
	Div-dp [78]	Medium	Top-k
	Div-cut [78]	High	Top-k

Algorithm with Monotone Submodular Set Functions [5] are suitable for all cases. DIVGEN Algorithm [3] also applies in top-k problem. Max-sum dispersion (MSD) [32] and Clustering-Based Method [93] are appropriate for graph problem.

In intent-based diversification, most frequent method [79] is of high efficiency, while efficiency of OptSelect [9] is high-to-medium. In addition, maximum result variety method [79] and Topic Diversification Algorithm [110] are of medium efficiency, while efficiency of Most Satisfied method [79] is low. OptSelect [9] and Topic Diversification Algorithm [110] are suitable for all cases. Most frequent method [79], maximum result variety method [79] and most satisfied method [79] apply in web search.

In addition, Div-cut [78] is high efficiency. Efficiency of Div-dp [78] is medium, while Div-astar [78] is of low efficiency.

7.2 Effectiveness issues

Some approaches are proposed to improve the effectiveness of diversification.

Drosou et al. [24] propose a novel and intuitive definition of diversity named DisC diversity. It aims to select a minimum representative subset Ω of a result set R, such that each object in R can be represented by a similar object in Ω, and the objects contained in Ω are dissimilar to each other. A radius, denoted as r, around each object is utilized to model the dissimilarity. Then such subsets are called r-DisC diverse subsets of R. A novel adaptive diversification method is presented through decreasing r, termed zooming-in, and increasing r, termed zooming-out. It is shown that locating r-DisC diverse subsets is an NP-hard problem. Therefore, some efficient heuristics with approximation are provided, including incremental heuristic algorithms for zooming-in and zooming-out, and corresponding theoretical bounds are also provided. Finally, an implementation based on spatial indexing is proposed and extensive experiments are conducted to verify these techniques' performance.

[109] converts diversification as a learning problem. The ranking function and loss function are proposed. With the consideration that the diverse ranking is a sequential selection process, the ranking function is defined as the combination of relevance score and diversity score between the current object and selected objects. The loss function is defined as the likelihood loss of ground truth based on Plackett–Luce model. Thus, the diverse ranking list is generated by a sequential selection process based on the ranking function. Following up [109], in [100], a perceptron-based learning algorithm is presented to learn the model as the combination of relevance and diversity from the training data. In such algorithm, positive and negative diverse rankings are constructed for each query. Then the parameters are adjusted to maximize the margins between the positive and negative rankings.

As stated in Sect. 2, the tradeoff parameter of relevance and diversity is difficult to generate. To address this issue, Yu et al. in [103] formulate the result diversification as a 0–1 multiple subtopic knapsack problem (MSKP) with absence of the tradeoff parameter. Such problem optimally chooses a subset of objects to fill up multiple subtopic knapsacks. Since it is proven to be a NP-hard problem, the max-sum belief propagation algorithm is proposed to solve it.

7.3 Efficiency issues

Various indices and algorithms are adopted to accelerate diversifying query results. Li et al. [64] present one and conduct a thorough study on it. First, a novel approach for evaluating diversity queries based on the concept of computing a core cover of a query is introduced. Then based on this concept, a new index method, denoted as D-Index, and two of its variants, namely, D-Tree as well as D^{+}-Tree, are designed to evaluate both static diversity queries and dynamic diversity queries.

Khan el al. [57] utilize partial distance computations to reduce CPU cost of diversification. That is, for high-dimensional data, the distance of a little share of dimensions between the candidate and the tuples in the results is computed. Thus, some hopeless tuples are pruned. New dimensions are added progressively. Such approach is suitable for vertically partitioned data stores. In such database engine, the query processing and diversification are integrated to achieve high performance.

7.4 Evaluation

7.4.1 Measures

precision and recall Precision and recall scores are traditional measures to evaluate experimental results. Drosou and Pitoura [22] and Zhang et al. [106] use precision and recall measures as follows.

$$\text{Redundancy—precision} = \frac{R^-}{R^- + N^-}$$
$$\text{Redundancy—recall} = \frac{R^-}{R^- + R^+}$$

where R^- is the set of non-delivered redundant documents, N^- is the set of non-delivered non-redundant ones and R^+ is the set of delivered redundant ones.

In Liu et al. [67,110], average recall scores and average precision scores are used to measure results. Carbonell and Goldstein [10], Kraaij et al. [59], Miller et al. [70] and Vieira et al. [95] use average precision scores as measure. Besides, [105] uses interpolated precision (PR curve), initial precision, non-interpolated average precision and recall measures.

For example, we assume that the filtering system identifies relevant documents with 100 % precision and recall by evaluating redundancy filtering only on a stream of documents marked relevant by NIST assessors. In some tests, we also evaluate the effectiveness of redundancy scoring algorithm and factor out the effect of the redundancy threshold algorithm, by reporting average Precision and Recall figures for redundant documents.

nDCG value In Ziegler et al. [1,110], NDCG-IA values are used to measure degree of diversification as follows.

$$\text{NDCG—IA}(S, k) = \sum_{c} P(c|q)\text{NDCG}(S, k|c)$$

[9,16,53,83] uses α-nDCG values measure. Hu et al. [53] also uses D♯-nDCG values as measure to evaluate.

For example, when documents are all of single intent, to maximize NDCG—IA, these documents should be ordered in decreasing level of relevance relative to each intent. NDCG—IA takes into account the distribution of intents, and forces a tradeoff between adding documents with higher relevance scores and those that cover additional intents. It can be interpreted as what the “average” user will think of the NDCG of the given ordering of the results.

diversity [79] uses measure of diversity as follows.

Given a set of results to rerank D.

$$\text{diversity}(D) = \max_{d \epsilon D} \text{match}(d, u).$$
$$\text{match}_{\text{unigram}}(d, u) = \sum_{t \epsilon d} w_t$$
$$w_t = \log \frac{(r_t + 0.5)(N - n_t + 0.5)}{(n_t + 0.5)(R - r_t + 0.5)}$$

where N is the number of documents in the corpus, R is the number for ones have relevance feedback, and n_t and r_t are the numbers of documents in N and R that contain the term t.

In Yu et al. [104] and Drosou and Pitoura [24], average of Jaccard diversity distance is used to measure diversification as follows:

$$\text{Jaccard}(S(o_i), S(o_j)) = 1 - \frac{|S(o_i) \cap S(o_j)|}{|S(o_i) \cup S(o_j)|}$$

[16] also uses the average rank of diversity as one of measures.

relevance Fractional difference in relevance is used as the measure in [32] as follows.

Given a diversified list $D(q)$ with the original list $R_k(q)$.

$$\mathrm{Rel}(s,q)=\sum_{d\varepsilon D(q)}\frac{1}{\mathrm{pos}(d)}p_q(d,s)$$

where s is a topic, pos(d) is the 1-based rank of d in the list $D(q)$.

$$\mathrm{FR}_q=\frac{\mathrm{Relevance}_q(D(q))-\mathrm{Relevance}_q(R_k(q))}{\max(\mathrm{Relevance}_q(D(q)),\mathrm{Relevance}_q(R_k(q)))}$$

Also, [104] uses relevance measure as follows:

$$\mathrm{relevance}(u,i)=\sum_{i'\epsilon I}\mathrm{ItemSim}\left(i,i'\right)\times\mathrm{rating}\left(u,i'\right)$$

where i is an item, ItemSim(i, i') returns a measure of similarity between two items i and i', and rating(u, i') indicates the rating of item i' by user u.

Many other measures can be used in diversification evaluation. For instance, [2,24,94] uses solution data size as the measure tool. Capannini et al. [9] and Santos et al. [83] use IA-P as the measure. Fractional novelty is used as measure in [32] as follows.

Given a list L. S_q a set of Wikipedia disambiguation pages for q. Each topic $s\ \epsilon\ S_q$. $p_q(x,s)$ means the confidence on the topic over all the results in L. $\mathbf{I}(\cdot)$ is the indicator function for an expression.

$$\mathrm{Novelty}_q(L)=\frac{1}{\left|S_q\right|}\sum_{s\epsilon S_q}\mathbf{I}\left(\sum_{x\epsilon L}p_q(x,s)>\theta\right)$$

Given two lists $D(q)$ and $R_k(q)$.

$$FN_q=\frac{\mathrm{Novelty}_q(D(q))-\mathrm{Novelty}_q(R_k(q))}{\max(\mathrm{Novelty}_q(D(q)),\mathrm{Novelty}_q(R_k(q)))}$$

[21] uses ranking function $\mathrm{f}_R(\mathrm{p}_i)$ that takes value in [0,1], while [2] evaluates via the most frequent categories of the search results. MAP-IA and MRR-IA values are used to weigh experimental results in [1]. Hu et al. [53] measures ERR-IA and NRBP values. Coverage and orthogonality of the summaries are both used in [67] as follows.

Given a summary S, we evaluate its coverage as the percentage of terms from T that appears in the sentences in S.

$$\mathrm{Coverage}(S)=\frac{\left|\bigcup_{s\epsilon S}s\right|}{|T|}.$$

$$\mathrm{Orthogonality}(S)=\frac{2}{|S|\,(|S|-1)}\sum_{s,s'\epsilon Ss\neq s'}J_D\left(s,s'\right).$$

where s and s' are two sentences, $J_D\left(s,s'\right)$ is Jaccard distance between s and s'.

For example, for any summary S, we have $0\leq\mathrm{Coverage}(S)\leq 1$. The closer the coverage value is to 1 the better the coverage of a summary. Similarly, for two sentences s and s', the value of $J_D\left(s,s'\right)$ takes values between 0 and 1, and so does Orthogonality(S). The closer the value of Orthogonality(S) is to 1, the more orthogonal the summary.

Three measures are used in [54], the average and worst case ratio of the result set scores, the average distance of the points in result set and the percentage of tuples read.

Also, two measures, Folwkes–Mallows index(FM) and the variation of information(VI), in [93] are as follows.

Given two clusterings C and C'.

$$\mathrm{FM}\left(C, C'\right) = \sqrt{W_R\left(C, C'\right) W_{RR}\left(C, C'\right)}$$

where W_R and $W_R R$ are two asymmetric criteria proposed by Wallance [97].

$$\mathrm{VI}\left(C, C'\right) = \left[H(C) - I\left(C, C'\right)\right] + \left[H\left(C'\right) - I\left(C, C'\right)\right]$$

where $H\left(C'\right)$ is the uncertainty about cluster in C' of a random image, $I\left(C, C'\right)$ is the sum of the corresponding entropies taken over all possible pairs of clusters.

7.4.2 Data sets

In order to provide a comprehensive evaluation of diversification, extensive experiments have been conducted on plenty of real and synthetic datasets.

Real Datasets

Datasets extracted from web pages Most of real datasets are extracted from web pages. Works and their corresponding real datasets are illustrated in Table 3.

Datasets from TREC text retrieval conference (TREC) provide a great deal of experimental data. Capannini et al. [9] and Santos et al. [83] use the subset of the TREC ClueWeb09 dataset [37]. Also, [16] conducts experiments on collections from the TREC 2005 and 2006 question answering tracks. TREC-6 test collection is used in [70], while TREC-7 SDR test collection is used in [59,70]. Besides, [106] conducts part of experiments on TREC Interactive Dataset. Most of experiments in [105] are done on TREC Disk 4 and 5 (minus Congressional Record) with topics 401–450 and TREC8 small web collection with topics 401–450.

Collected Datasets It is also an effective way to collect data from people. For instance, in Radlinski and Dumais [79], data are gathered from 33 volunteers. Similarly, [10] uses data collected from five users who were undergraduates from various disciplines.

Synthetic Datasets Some experiments are performed on synthetic datasets. In [21], synthetic datasets consisting of 200 data points in the euclidean space. Liu et al. [67] uses synthetic datasets ClusteredDUC and MixedDUC created based on the 2002 Document Understanding Conference (DUC) [99]. Besides, [5] conducts experiments on synthetic data which is generated uniformly at random within a given range. Also, part of experiments in [24] are done on objects are either uniformly distributed in space (“Uniform”) or form (hyper) spherical clusters of different sizes (“Clustered”).

8 Research challenges

Even though many approaches have been proposed, many facets of query result diversification are not studied thoroughly or even paid attention to. Several challenging directions which are worth further exploring are discussed in this section.

The concept of diversity is useful for increasing the usability of results, and improving the experience of users. Even though diversification has been applied to various kinds of queries and mining tasks as surveyed in this paper, it is not studied in many kinds of query processing and data mining tasks. One important kinds of task is the similarity search, especially the similarity search based on query relaxation. Query relaxation has been proposed to increase the number of results when too few results are returned. When relaxation and diversity are combined, the quality of results should be improved furthermore. Another benefit of applying diversity on similarity search is to overcome the difficulty of selection the similarity threshold.

Table 3 Datasets extracted from web pages

Work	Datasets
Max-sum dispersion [32]	[39,47]
Query result diversification [22]	[46]
Topic diversification [110]	
Computation of diverse query results [94]	Car listings from Yahoo! autos
Swap method [104]	del.icio.us [38]
BSwap method [104]	Y!Movies [44]
Maximal marginal relevance (MMR) method [104]	
Publish/subscribe delivery [27]	Movies [41]
Search result sampling [2]	A set of 1.8 million Web pages (1.1 billion words)
Search result diversification [1]	10,000 Random queries returned from the three commercial search engines
Clustering-based method [93]	75 Topics that were randomly selected from the Flickr search logs
Diversification with hierarchical intents [53]	Datasets provided by the IMine (Intent Mining) task in NTCIR-11 [68]
Divdb system [96]	ALOI [30]
	Faces [42]
Motley method [54]	Forest cover dataset [40]
Greedy marginal contribution (MMC) method [95]	Faces [50]
Greedy randomized with neighborhood expansion (GNE) method [95]	Nasa [48]
	Colors [43]
	Dblp [49]
	Docs [45]
DisC diversity [24]	Acme digital cameras database [36]
	Greek cities dataset [51]

Instead, the results that are similar to the query and diverse enough are returned. Thus, it is worth to study the application of diversity with similarity search for strings, relational tuples, twigs in XML or subgraphs. However, the challenges include the measure of result quality as the combination of diversity and similarity as well as the development of efficient algorithms suitable for massive data.

For practical issues of results diversification, distribute platform such as Mapduce and Spark should be used to satisfy the requirements of scalability. Since existing concepts of diversification are proposed as the global feature of result set, the challenge of distributed query processing with the consideration of result diversification is to define the diversification of the partial results on distributed machines and the combination strategy for partial results retrieved from separate machines to generate a uniform diverse result set. Moreover, with the consideration of information integration from heterogenous data sources, the result diversification could still be considered. The challenges brought by this issue include the evaluation of result diversification in data sources, data source selection based on result diversification, query partition and translation with the consideration of result diversification as well as the combination of partial results returned from data sources to maximize the diversification.

Apart from existing definitions, after collecting various sources of information, the concept of freshness should be considered. However, in early researches, freshness is considered the same as novelty [73,90,91], where novelty stands for documents containing fresh contents and the objective is to improve novelty. Nonetheless, these two concepts have differences. Novelty mainly describes the new information from objects selected comparing to previously objects, but freshness lays particular emphasis on the content change of autonomous data sources (especially the Web). The concept of freshness of a content fragment and that of a dynamic Web page are firstly introduced in [77]. When studying query result diversification, we fail to take into consideration the autonomous change in the data sources. Thus, some fresh information is neglected and cannot be offered to users in time. To bridge this problem, freshness can be taken into account. In [6,15,56], several techniques to improve freshness are proposed in general databases, information systems, and real-time databases respectively. Then in [60,61], how to balance performance and freshness with regard to different scenarios is studied. In this sense, in order to further improve users' satisfaction, improving freshness can be a future research direction. The challenge of this direction is to propose new measure for freshness and diversity as well as query processing techniques.

From the aspect of data quality, diversity could be treated as one dimension of result quality. It is still a problem to combine it with other dimensions of data quality such as completeness and consistency to define a complete quality of result set. Diversity is not orthogonal with other data quality dimensions. The relationship between diversity and other data quality dimensions is left for an open problem. Thus, the challenges of the measure of result set quality with the consideration of diversification and efficiently query processing algorithm development with the complete result set quality measure.

Last but not the least, from our survey and study, we note that content-based diversification, intent-based diversification and novelty-based diversification are mostly modeled and analyzed separately to improve users' satisfaction. However, these three concepts are not independent from each other. They share some overlapped features and could be combined to satisfy the users' needs. In this sense, it is beneficial to combine these three kinds of diversification with several tradeoff parameters and establish a complete methodology for diversification to achieve more reasonable diversity. The major challenge is to develop a general optimization goal for optimization and the efficient algorithms with the optimization goal.

These open research directions have technical challenges and are well worth exploring to a further and deeper extent.

9 Conclusions

In the last few years, as users' information needs are increasingly becoming more extensive, many information systems are laying more emphases on improving users' satisfaction. Generally, users' satisfaction toward a query is not only decided by the relevance between the retrieved results and the query, but also related to the diversity of results. This is because various retrieved information can arouse users' interest as well as provide more information sources for them. Therefore, query result diversification is very meaningful and important.

This paper is a survey of query result diversification, which provides the definitions and various existing algorithms of three kinds of methods to diversify the retrieved results, namely content-based, intent-based, and novelty-based diversification. Then it also gives a summarization of diversification in some specialized search areas, such as web search, time series search, and so on. Moreover, this survey discusses the novel indexes, techniques, tools as well as systems to show that diversification methods can be widely applied in practice. Finally, this paper offers some ideas about future research directions which have not been thoroughly studied or paid enough attention to and then inspires researchers to explore these areas.

Acknowledgements This paper was partially supported by National Sci-Tech Support Plan 2015BAH10F01 and NSFC Grant U1509216, 61472099, 61133002 and the Scientific Research Foundation for the Returned Overseas Chinese Scholars of Heilongjiang Province LC2016026.

References

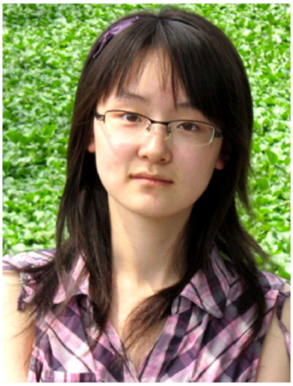
1. Agrawal R, Gollapudi S, Halverson A, Ieong S (2009) Diversifying search results. In: Proceedings of the 2nd ACM international conference on web search and data mining. ACM, pp 5–14
2. Anagnostopoulos A, Broder AZ, Carmel D (2006) Sampling search-engine results. World Wide Web 9(4):397–429
3. Angel A, Koudas N (2011) Efficient diversity-aware search. In: Proceedings of the 2011 ACM SIGMOD international conference on management of data. ACM, pp 781–792
4. Berchtold S, Ertl B, Keim DA, Kriegel H-P, Seidl T (1998) Fast nearest neighbor search in high-dimensional space. In: 1998. Proceedings, 14th international conference on data engineering. IEEE, pp 209–218
5. Borodin A, Lee HC, Ye Y (2012) Max-sum diversification, monotone submodular functions and dynamic updates. In: Proceedings of the 31st symposium on principles of database systems. ACM, pp 155–166
6. Bouzeghoub M (2004) A framework for analysis of data freshness. In: Proceedings of the 2004 international workshop on information quality in information systems. ACM, pp 59–67
7. Broder AZ, Charikar M, Frieze AM, Mitzenmacher M (1998) Min-wise independent permutations. In: Proceedings of the 30th annual ACM symposium on theory of computing. ACM, pp 327–336
8. Cao X, Chen L, Cong G, Jensen CS, Qu Q, Skovsgaard A, Wu D, Yiu ML (2012) Spatial keyword querying. In: Atzeni P , Cheung D , Ram S (eds) Conceptual modeling. Springer, Berlin, pp 16–29
9. Capannini G, Nardini FM, Perego R, Silvestri F (2011) Efficient diversification of web search results. Proc VLDB Endow 4(7):451–459
10. Carbonell J, Goldstein J (1998) The use of mmr, diversity-based reranking for reordering documents and producing summaries. In: Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval. ACM, pp 335–336
11. Catallo I, Ciceri E, Fraternali P, Martinenghi D, Tagliasacchi M (2013) Top-*k* diversity queries over bounded regions. ACM Trans Database Syst (TODS) 38(2):10
12. Chakraborty T, Modani N, Narayanam R, Nagar S (2015) Discern: a diversified citation recommendation system for scientific queries. In: 31st IEEE international conference on data engineering, ICDE 2015, Seoul, Apr 13–17, pp 555–566
13. Chen L, Cong G (2015) Diversity-aware top-*k* publish/subscribe for text stream. In: Proceedings of the 2015 ACM SIGMOD international conference on management of data, Melbourne, May 31–June 4, pp 347–362
14. Chen Y-Y, Suel T, Markowetz A (2006) Efficient query processing in geographic web search engines. In: Proceedings of the 2006 ACM SIGMOD international conference on management of data. ACM, pp 277–288
15. Cho J, Garcia-Molina H (2000) Synchronizing a database to improve freshness. In: Cho J, Garcia-Molina H (eds) ACM sigmod record, vol 29(2). ACM, pp 117–128
16. Clarke CL, Kolla M, Cormack GV, Vechtomova O, Ashkan A, Büttcher S, MacKinnon I (2008) Novelty and diversity in information retrieval evaluation. In: Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval. ACM, pp 659–666
17. Demidova E, Fankhauser P, Zhou X, Nejdl W (2010) Divq: diversification for keyword search over structured databases. In: Proceedings of the 33rd international ACM SIGIR conference on research and development in information retrieval. ACM, pp 331–338
18. Deng T, Fan W (2013) On the complexity of query result diversification. ACM Trans Database Syst (TODS) 6(2):577–588

19. Deshpande M, Karypis G (2004) Item-based top-*n* recommendation algorithms. ACM Trans Inf Syst (TOIS) 22(1):143–177
20. Dou Z, Hu S, Chen K, Song R, Wen J-R (2011) Multi-dimensional search result diversification. In: Proceedings of the 4th ACM international conference on web search and data mining. ACM, pp 475–484
21. Drosou M, Pitoura E (2009) Diversity over continuous data. IEEE Data Eng Bull 32(4):49–56
22. Drosou M, Pitoura E (2010) Search result diversification. ACM SIGMOD Rec 39(1):41–47
23. Drosou M, Pitoura E (2012) Dynamic diversification of continuous data. In: Proceedings of the 15th international conference on extending database technology. ACM, pp 216–227
24. Drosou M, Pitoura E (2012) Disc diversity: result diversification based on dissimilarity and coverage. Proc VLDB Endow 6(1):13–24
25. Drosou M, Pitoura E (2013) Poikilo: a tool for evaluating the results of diversification models and algorithms. Proc VLDB Endow 6(12):1246–1249
26. Drosou M, Pitoura E (2014) Diverse set selection over dynamic data. IEEE Trans Knowl Data Eng 26(5):1102–1116
27. Drosou M, Stefanidis K, Pitoura E (2009) Preference-aware publish/subscribe delivery with diversity. In: Proceedings of the 3rd ACM international conference on distributed event-based systems. ACM, p 6
28. Eravci B, Ferhatosmanoglu H (2013) Diversity based relevance feedback for time series search. Proc VLDB Endow 7(2):109–120
29. Fan W, Wang X, Wu Y (2013) Diversified top-*k* graph pattern matching. Proc VLDB Endow 6(13):1510–1521
30. Geusebroek J, Burghouts GJ, Smeulders AWM (2005) The amsterdam library of object images. Int J Comput Vis 61(1):103–112
31. Gollapudi S, Panigrahy R (2006) Exploiting asymmetry in hierarchical topic extraction. In: Proceedings of the 15th ACM international conference on information and knowledge management. ACM, pp 475–482
32. Gollapudi S, Sharma A (2009) An axiomatic approach for result diversification. In: Proceedings of the 18th international conference on world wide web. ACM, pp 381–390
33. Gonzalez TF (1985) Clustering to minimize the maximum intercluster distance. Theoret Comput Sci 38:293–306
34. Haritsa JR (2009) The kndn problem: a quest for unity in diversity. IEEE Data Eng Bull 32(4):15–22
35. Herlocker JL, Konstan JA, Borchers A, Riedl J (1999) An algorithmic framework for performing collaborative filtering. In: Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval. ACM, pp 230–237
36. http://acme.com/digicams
37. http://boston.lti.cs.cmu.edu/Data/clueweb09/
38. http://del.icio.us/
39. http://en.wikipedia.org/wiki/Disambiguation_page
40. http://ftp.ics.uci.edu/pub/machine-learning-databases/covtype
41. http://had.co.nz/data/movies
42. http://iam.unibe.ch/pub/Images/FaceImages
43. http://kdd.ics.uci.edu
44. http://movies.yahoo.com/
45. http://trec.nist.gov
46. http://www.bookcrossing.com
47. http://www.csie.ntu.edu.tw/~cjlin/liblinear/
48. http://www.dimacs.rutgers.edu/Challenges/Sixth/software.html
49. http://www.informatik.uni-trier.de/~ley/db
50. http://www.informedia.cs.cmu.edu
51. http://www.rtreeportal.org
52. Huang X, Cheng H, Li R-H, Qin L, Yu JX (2013) Top-*k* structural diversity search in large networks. Proc VLDB Endow 6(13):1618–1629
53. Hu S, Dou Z, Wang X, Sakai T, Wen J (2015) Search result diversification based on hierarchical intents. In: Proceedings of the 24th ACM international on conference on information and knowledge management, CIKM 2015, Melbourne, Oct 19–23, pp 63–72
54. Jain A, Sarda P, Haritsa JR (2004) Providing diversity in *k*-nearest neighbor query results. In: Dai H, Srikant R, Zhang, C (eds) Advances in knowledge discovery and data mining. Springer, Berlin, pp 404–413
55. Jones WP, Furnas GW (1987) Pictures of relevance: a geometric analysis of similarity measures. J Am Soc Inf Sci 38(6):420–442

56. Kang K-D, Son SH, Stankovic JA, Abdelzaher TF (2002) A qos-sensitive approach for timeliness and freshness guarantees in real-time databases. In: 2002. Proceedings. 14th Euromicro conference on real-time systems. IEEE, pp 203–212
57. Khan HA, Sharaf MA (2015) Progressive diversification for column-based data exploration platforms. In: 31st IEEE international conference on data engineering, ICDE 2015, Seoul, Apr 13–17, pp 327–338
58. Khan HA, Drosou M, Sharaf MA (2013) Dos: an efficient scheme for the diversification of multiple search results. In: International conference on scientific and statistical database management, pp 1–4
59. Kraaij W, Pohlmann R, Hiemstra D (2000) Twenty-one at trec-8: using language technology for information retrieval. In: Voorhees E, Harman D (eds) National institute of standards and technology
60. Labrinidis A, Roussopoulos N (2003) Balancing performance and data freshness in web database servers. In: Proceedings of the 29th international conference on very large data bases vol 29. VLDB Endowment, pp 393–404
61. Labrinidis A, Roussopoulos N (2004) Exploring the tradeoff between performance and data freshness in database-driven web servers. VLDB J 13(3):240–255
62. Lafferty J, Zhai C (2003) Probabilistic relevance models based on document and query generation. In: Croft W B, Lafferty J (eds) Language modeling for information retrieval. Springer, Netherlands, pp 1–10
63. Lee L (1999) Measures of distributional similarity. In: Proceedings of the 37th annual meeting of the association for computational linguistics on computational linguistics. Association for Computational Linguistics, pp 25–32
64. Li L, Chan C-Y (2013) Efficient indexing for diverse query results. Proc VLDB Endow 6(9):745–756
65. Li R-H, Yu JX (2013) Scalable diversified ranking on large graphs. IEEE Trans Knowl Data Eng 25(9):2133–2146
66. Liu Z, Sun P, Chen Y (2009) Structured search result differentiation. Proc VLDB Endow 2(1):313–324
67. Liu K, Terzi E, Grandison T (2009) Highlighting diverse concepts in documents. In: Proceedings of the SIAM International Conference on Datamining (SDM), pp 545–556
68. Liu Y, Song R, Zhang M, Dou Z, Yamamoto T, Kato MP, Ohshima H, Zhou K (2014) Overview of the NTCIR-11 imine task. In: Proceedings of the 11th NTCIR conference on evaluation of information access technologies, NTCIR-11, National Center of Sciences, Tokyo, Dec 9–12
69. Martinenghi D, Tagliasacchi M (2010) Proximity rank join. Proc VLDB Endow 3(1–2):352–363
70. Miller DR, Leek T, Schwartz RM (1999) A hidden markov model information retrieval system. In: Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval. ACM, pp 214–221
71. Minack E, Siberski W, Nejdl W (2011) Incremental diversification for very large sets: a streaming-based approach. In: Proceedings of the 34th international ACM SIGIR conference on research and development in information retrieval. ACM, pp 585–594
72. Nanongkai D, Sarma AD, Lall A, Lipton RJ, Xu J (2010) Regret-minimizing representative databases. Proc VLDB Endow 3(1–2):1114–1124
73. Ng KW, Tsai FS, Chen L, Goh KC (2007) Novelty detection for text documents using named entity recognition. In: 2007 6th international conference on information, communications & signal processing. IEEE, pp 1–5
74. Ni J, Ravishankar CV (2007) Pointwise-dense region queries in spatio-temporal databases. In 2007. ICDE 2007. IEEE 23rd international conference on data engineering. IEEE, pp 1066–1075
75. Ntoutsi E, Stefanidis K, Rausch K, Kriegel H (2014) "Strength lies in differences": diversifying friends for recommendations through subspace clustering. In: Proceedings of the 23rd ACM international conference on conference on information and knowledge management, CIKM 2014, Shanghai, Nov 3–7, pp 729–738
76. Ozdemiray AM, Altingovde IS (2014) Query performance prediction for aspect weighting in search result diversification. In: Proceedings of the 23rd ACM international conference on conference on information and knowledge management, CIKM 2014, Shanghai, Nov 3–7, pp 1871–1874
77. Papastavrou S, Chrysanthis PK, Samaras G (2013) Performance vs. freshness in web database applications. In: World wide web, pp 1–27
78. Qin L, Yu JX, Chang L (2012) Diversifying top-*k* results. Proc VLDB Endow 5(11):1124–1135
79. Radlinski F, Dumais S (2006) Improving personalized web search using result diversification. In: Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval. ACM, pp 691–692
80. Rafiei D, Bharat K, Shukla A (2010) Diversifying web search results. In: Proceedings of the 19th international conference on world wide web. ACM, pp 781–790
81. Resnick P, Iacovou N, Suchak M, Bergstrom P, Riedl J (1994) Grouplens: an open architecture for collaborative filtering of netnews. In: Proceedings of the 1994 ACM conference on computer supported cooperative work. ACM, pp 175–186

82. Robertson SE (1977) The probability ranking principle in IR. J Doc 33(4):294–304
83. Santos RL, Macdonald C, Ounis I (2010) Exploiting query reformulations for web search result diversification. In: Proceedings of the 19th international conference on world wide web. ACM, pp 881–890
84. Santos LFD, Oliveira WD, Ferreira MRP, Traina AJM, Traina C (2013) Parameter-free and domain-independent similarity search with diversity. In: International conference on scientific and statistical database management, pp 1–12
85. Sarwar B, Karypis G, Konstan J, Riedl J (2001) Item-based collaborative filtering recommendation algorithms. In: Proceedings of the 10th international conference on world wide web. ACM, pp 285–295
86. Schedl M, Hauger D (2015) Tailoring music recommendations to users by considering diversity, mainstreaminess, and novelty. In Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval, Santiago, Aug 9–13, pp 947–950
87. Stefanidis K, Drosou M, Pitoura E (2010) Perk: personalized keyword search in relational databases through preferences. In Proceedings of the 13th international conference on extending database technology. ACM, pp 585–596
88. Tang J, Sanderson M (2010) Evaluation and user preference study on spatial diversity. In: Gurrin C, He Y, Kazai G, Kruschwitz U, Little S, Roelleke T, Ruger S, Rijsbergen KV (eds) Advances in information retrieval. Springer, Berlin, pp 179–190
89. Tong H, He J, Wen Z, Konuru R, Lin C-Y (2011) Diversified ranking on large graphs: an optimization viewpoint. In: Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 1028–1036
90. Tsai FS, Chan KL (2010) Redundancy and novelty mining in the business blogosphere. Learn Organ 17(6):490–499
91. Tsai FS, Kwee AT (2011) Database optimization for novelty mining of business blogs. Expert Syst Appl 38(9):11040–11047
92. Van Kreveld M, Reinbacher I, Arampatzis A, Van Zwol R (2005) Multi-dimensional scattered ranking methods for geographic information retrieval*. GeoInformatica 9(1):61–84
93. van Leuken RH, Garcia L, Olivares X, van Zwol R (2009) Visual diversification of image search results. In: Proceedings of the 18th international conference on world wide web. ACM, pp 341–350
94. Vee E, Srivastava U, Shanmugasundaram J, Bhat P, Yahia SA (2008) Efficient computation of diverse query results. In 2008. ICDE 2008. IEEE 24th international conference on data engineering. IEEE, pp 228–236
95. Vieira MR, Razente HL, Barioni MCN, Hadjieleftheriou M, Srivastava D, Traina A, Tsotras VJ (2011) On query result diversification. In: 2011 IEEE 27th international conference on data engineering (ICDE). IEEE, pp 1163–1174
96. Vieira MR, Razente HL, Barioni MC, Hadjieleftheriou M, Srivastava D, Traina C Jr, Tsotras VJ (2011) Divdb: a system for diversifying query results. Proc VLDB Endow 4(12):1395–1398
97. Wallace DL (1983) Comment. J Am Stat Assoc 78(383):569–576
98. Wang J, Cheng J, Fu AW-C (2013) Redundancy-aware maximal cliques. In: Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 122–130
99. www.nlpir.nist.gov/projects/duc/guidelines/2002.html
100. Xia L, Xu J, Lan Y, Guo J, Cheng X (2015) Learning maximal marginal relevance model via directly optimizing diversity evaluation measures. In: Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval, Santiago, Aug 9–13, pp 113–122
101. Xin D, Cheng H, Yan X, Han J (2006) Extracting redundancy-aware top-*k* patterns. In: Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 444–453
102. Yuan L, Qin L, Lin X, Chang L, Zhang W (2015) Diversified top-*k* clique search. In: 31st IEEE international conference on data engineering, ICDE 2015, Seoul, Apr 13–17, pp 387–398
103. Yu H, Ren F (2014) Search result diversification via filling up multiple knapsacks. In: Proceedings of the 23rd ACM international conference on conference on information and knowledge management, CIKM 2014, Shanghai, Nov 3–7, p 609–618
104. Yu C, Lakshmanan L, Amer-Yahia S (2009) It takes variety to make a world: diversification in recommender systems. In: Proceedings of the 12th international conference on extending database technology: advances in database technology. ACM, pp 368–378
105. Zhai C, Lafferty J (2001) Model-based feedback in the language modeling approach to information retrieval. In: Proceedings of the 10th international conference on information and knowledge management. ACM, pp 403–410
106. Zhang Y, Callan J, Minka T (2002) Novelty and redundancy detection in adaptive filtering. In: Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval. ACM, pp 81–88

107. Zhao F, Zhang X, Tung AK, Chen G (2011) Broad: Diversified keyword search in databases. Proc VLDB Endow 4(12):1355–1358
108. Zhu Y, Yu JX, Cheng H, Qin L (2012) Graph classification: a diversified discriminative feature selection approach. In: Proceedings of the 21st ACM international conference on information and knowledge management. ACM, pp 205–214
109. Zhu Y, Lan Y, Guo J, Cheng X, Niu S (2014) Learning for search result diversification. In: The 37th international ACM SIGIR conference on research and development in information retrieval, SIGIR '14, Gold Coast, July 06–11, p 293–302
110. Ziegler C-N, McNee SM, Konstan JA, Lausen G (2005) Improving recommendation lists through topic diversification. In: Proceedings of the 14th international conference on world wide web. ACM, pp 22–32

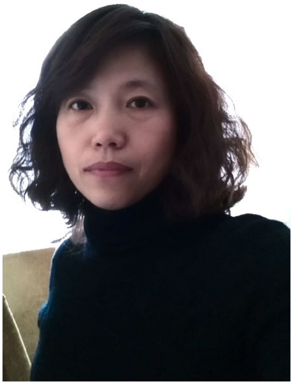


Kaiping Zheng born in 1992, female, is an undergraduate student majoring in Computer Science and Technology in Harbin Institute of Technology. Her research area is big data computing.

Hongzhi Wang born in 1978, male, PHD. He is a professor and doctoral supervisor at Harbin Institute of Technology. His research area is data management, including data quality, XML data management and graph management. He has published more than 100 papers in refereed journals and conferences. He is a recipient of the outstanding dissertation award of CCF, Microsoft Fellow and IBM PhD Fellowship.

Zhixin Qi born in 1994, female, is a graduate student majoring in Computer Science and Technology in Harbin Institute of Technology. Her research area is data quality.

Jianzhong Li was born in 1950. He is a professor and doctoral supervisor at Harbin Institute of Technology. He is a senior member of CCF. His research interests include database, parallel computing and wireless sensor networks, etc.



Hong Gao was born in 1966. She is a professor and doctoral supervisor at Harbin Institute of Technology. She is a senior member of CCF. Her research interests include data management, wireless sensor networks and graph database, etc.