Neural Comput & Applic
DOI 10.1007/s00521-015-1945-5

REVIEW

A survey on search results diversification techniques

Adnan Abid[1] · Naveed Hussain[2] · Kamran Abid[3] · Farooq Ahmad[5] · Muhammad Shoaib Farooq[1] · Uzma Farooq[1] · Sher Afzal Khan[4] · Yaser Daanial Khan[1] · Muhammad Azhar Naeem[3] · Nabeel Sabir[1]

Received: 9 July 2014 / Accepted: 5 June 2015

Abstract The quantity of information placed on the web has been greater than before and is increasing rapidly day by day. Searching through the huge amount of data and finding the most relevant and useful result set involves searching, ranking, and presenting the results. Most of the users probe into the top few results and neglect the rest. In order to increase user's satisfaction, the presented result set should not only be relevant to the search topic, but should also present a variety of perspectives, that is, the results should be different from one another. The effectiveness of web search and the satisfaction of users can be enhanced through providing various results of a search query in a certain order of relevance and concern. The technique used to avoid presenting similar, though relevant, results to the user is known as a diversification of search results. This article presents a survey of the approaches used for search result diversification. To this end, this article not only provides a technical survey of existing diversification techniques, but also presents a taxonomy of diversification algorithms with respect to the types of search queries.

Keywords Web search · Diversification of search results · Diversification techniques

1 Introduction

The Internet has converted into the main source of information, and web search appears as the main method for finding required information on the Internet. Search engines typically deliver an extended list of results that contains too many results, where relevant results tend to be alike. However, in order to make the result set informative as well as to increase the users' satisfaction, the search engines should not only present relevant results but should also present them in a diversified manner. Here, diversification can be defined by presenting the results to the user who covers all possible meanings of the input query, or to avoid presenting the same or similar, though relevant, results to the user again and again. Hence, from a user's perspective, the effectiveness of the presented search

✉ Yaser Daanial Khan
yaser.khan@umt.edu.pk

Adnan Abid
adnan.abid@umt.edu.pk

Naveed Hussain
naveed-hussain@usa.edu.pk

Kamran Abid
kamran@pu.edu.pk

Farooq Ahmad
FarooqAhmad@ciitlahore.edu.pk

Muhammad Shoaib Farooq
shoaib.farooq@umt.edu.pk

Uzma Farooq
uzma.farooq@umt.edu.pk

Muhammad Azhar Naeem
azhar@pu.edu.pk

Nabeel Sabir
nabeel.bloch@umt.edu.pk

[1] Department of Computer Science, University of Management and Technology, Lahore, Pakistan

[2] University of South Asia, Lahore, Pakistan

[3] University of the Punjab, Lahore, Pakistan

[4] Faculty of Computing and Information Technology in Rabigh, King Abdul Aziz University, Jeddah, Saudi Arabia

[5] COMSATS Institute of Information Technology, Lahore, Pakistan

Published online: 01 July 2015

results is first assessed in terms of relevance and then in terms of diversity.

Overall framework for search results diversification comprises of three main attributes: a relevance measure, diversity measure, and diversification objective. The relevance measure helps computing similarity of the search results to the input query; diversity measure helps identifying the novelty of each new result from the list of relevant results, whereas diversification objective defines the trade-off between the relevance and novelty of information to set up final ranking of diversified results.

Figure 1 presents the sequence of steps that define the search process, starting from taking an input query to presenting the results to the user. It shows that many steps are involved in the processing of a search query to obtain final diversified result set. A search query is passed to a search engine; the search engine in turn refines the query and uses an appropriate method to obtain relevant results; a certain number of top relevant results are selected to apply diversification; then a diversification method is applied only on the selected relevant results; and finally, relevant but diversified results are presented to the user.

1.1 Example 1.1

Consider a user *who plans a trip to a city and wants to find a hotel, a restaurant, and a cinema*. Figure 2 gives a result set of the user query. A simple way to answer the search query described here is to retrieve only top *k* relevant results on the first page. Figure 2a shows the results of this query while using relevance as the only criterion to rank the results for the user, a simple approach is to display only the relevant result of records with greater relevance score at the top of the first page. Here, it can be seen that the first three records are similar for hotel and restaurant, which

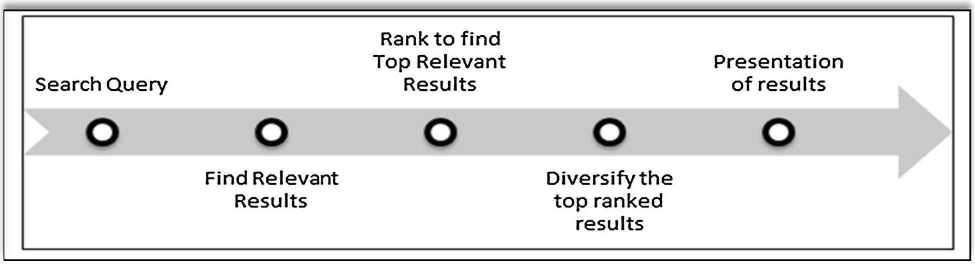
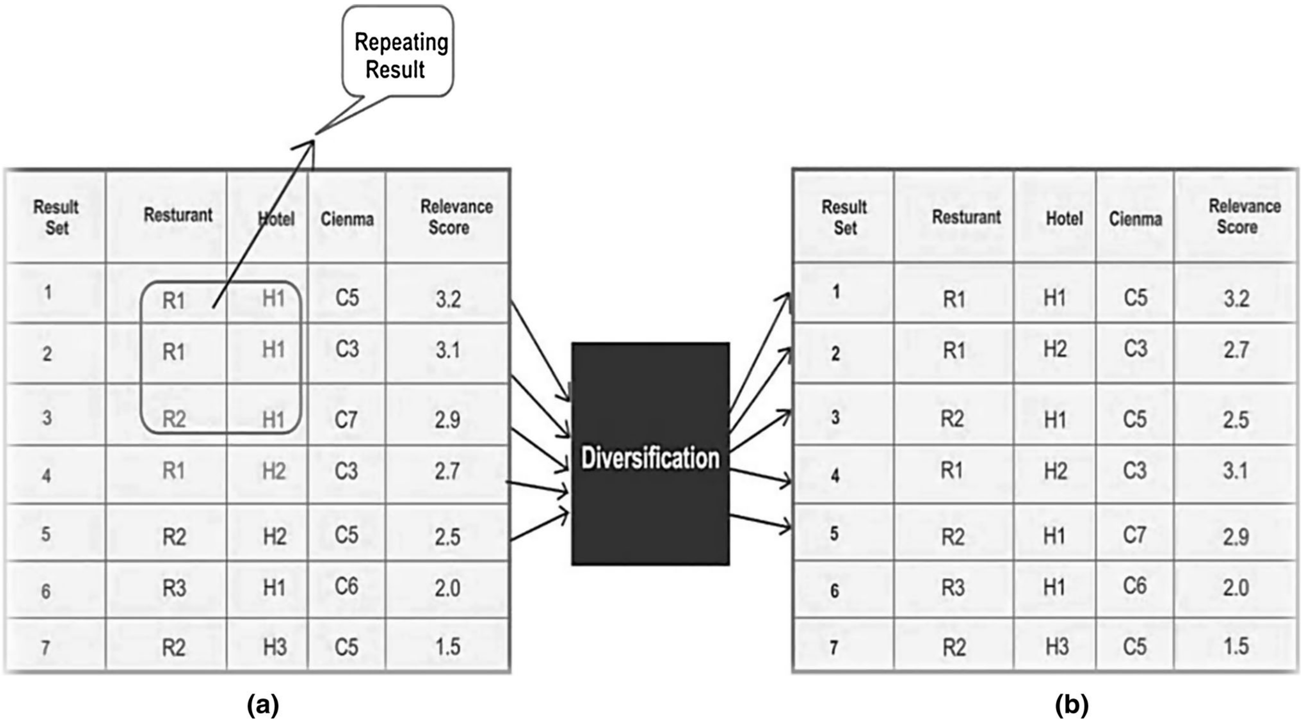

Search Query
Find Relevant Results
Rank to find Top Relevant Results
Diversify the top ranked results
Presentation of results

Fig. 1 Sequence of steps from search query to its final diversified result


Repeating Result
Diversification
(a)
(b)

Result Set	Resturant	Hotel	Cienma	Relevance Score
1	R1	H1	C5	3.2
2	R1	H1	C3	3.1
3	R2	H1	C7	2.9
4	R1	H2	C3	2.7
5	R2	H2	C5	2.5
6	R3	H1	C6	2.0
7	R2	H3	C5	1.5

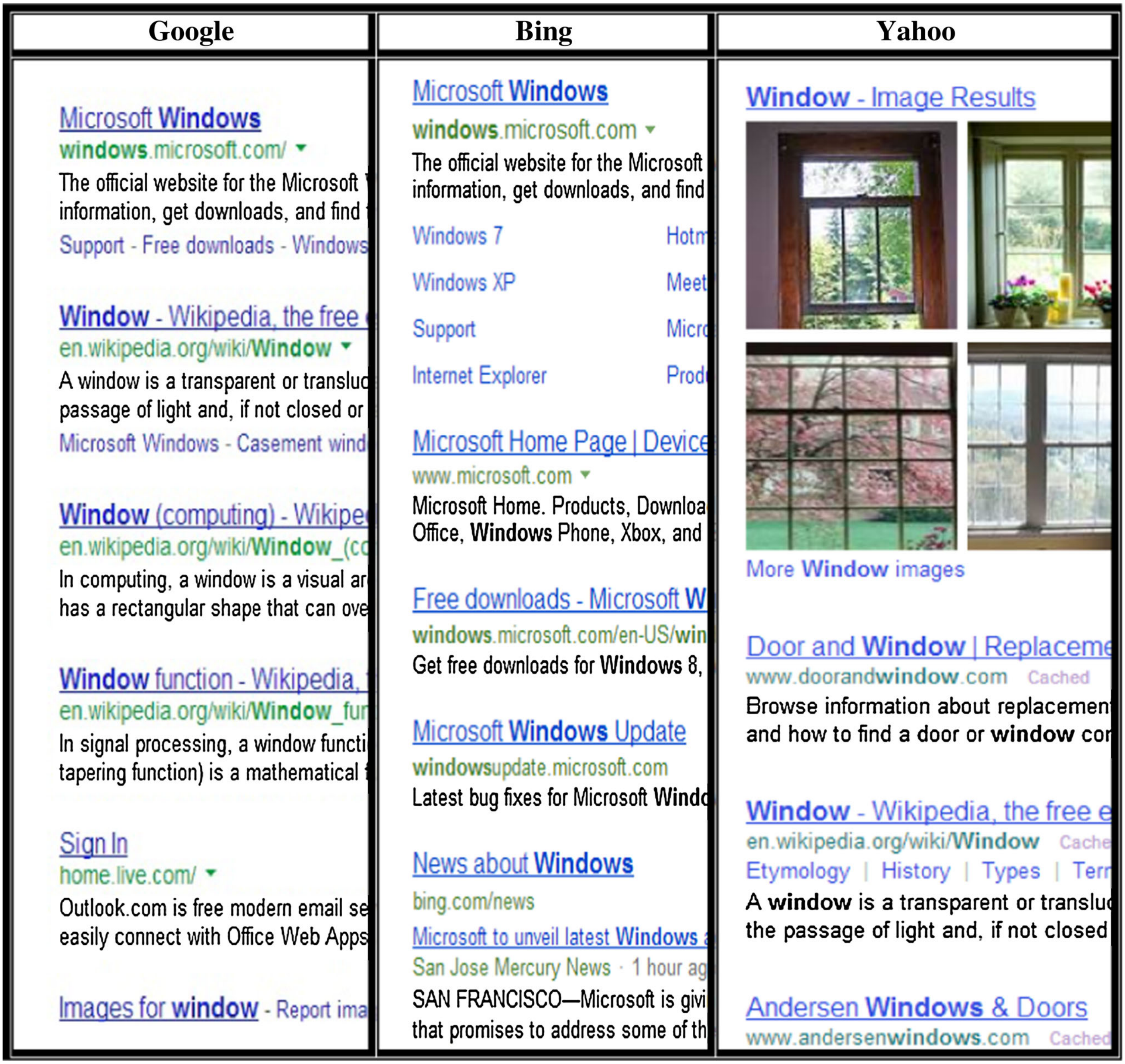
Result Set	Resturant	Hotel	Cienma	Relevance Score
1	R1	H1	C5	3.2
2	R1	H2	C3	2.7
3	R2	H1	C5	2.5
4	R1	H2	C3	3.1
5	R2	H1	C7	2.9
6	R3	H1	C6	2.0
7	R2	H3	C5	1.5

Fig. 2 A relevant and diversified result set of example query. **a** Relevant result. **b** Relevant + diversified results

makes the result set monotonous and boring for the user, and this is the problem with the search process that only involves relevance, but does not involve diversification. As a solution to the above problem, Fig. 2b shows the same set of results but in a different manner, by applying diversification to the results. This diversification technique re-ranks these search results to introduce diversity. Although this technique has minor compromise on relevance, yet scores of the results shown in the Fig. 2b present the results to the user in more satisfactory manner. Thus, diversification provides convenience to the user by providing search results from different perspectives.

1.2 Example 1.2

Consider the common single-term query "window." The user may relate term "window" to the Microsoft Windows operating system, or to the simple window that fits in house/office wall. These multiple possibilities, without providing any further information, make this query ambiguous. The modern search engines tend to create a set of results that cover different possible aspects of the input query. In Fig. 3 by using query "window," there is the result set of the three famous search engines. The figure shows that Google, Bing, and Yahoo order search



Google

Microsoft Windows
windows.microsoft.com/ ▾
The official website for the Microsoft
information, get downloads, and find
Support - Free downloads - Windows

Window - Wikipedia, the free
en.wikipedia.org/wiki/Window ▾
A window is a transparent or translu
passage of light and, if not closed or
Microsoft Windows - Casement wind

Window (computing) - Wikipe
en.wikipedia.org/wiki/Window_(cc
In computing, a window is a visual ar
has a rectangular shape that can ove

Window function - Wikipedia,
en.wikipedia.org/wiki/Window_fur
In signal processing, a window functi
tapering function) is a mathematical

Sign In
home.live.com/ ▾
Outlook.com is free modern email se
easily connect with Office Web Apps

Images for window - Report ima

Bing

Microsoft Windows
windows.microsoft.com ▾
The official website for the Microsoft
information, get downloads, and find

Windows 7 Hotm
Windows XP Meet
Support Micro
Internet Explorer Prod

Microsoft Home Page | Device
www.microsoft.com ▾
Microsoft Home. Products, Downloa
Office, Windows Phone, Xbox, and

Free downloads - Microsoft W
windows.microsoft.com/en-US/win
Get free downloads for Windows 8,

Microsoft Windows Update
windowsupdate.microsoft.com
Latest bug fixes for Microsoft Windo

News about Windows
bing.com/news
Microsoft to unveil latest Windows
San Jose Mercury News · 1 hour ag
SAN FRANCISCO—Microsoft is givi
that promises to address some of th

Yahoo

Window - Image Results
More Window images

Door and Window | Replaceme
www.doorandwindow.com Cached
Browse information about replacemen
and how to find a door or window cor

Window - Wikipedia, the free e
en.wikipedia.org/wiki/Window Cache
Etymology | History | Types | Terr
A window is a transparent or translu
the passage of light and, if not closed

Andersen Windows & Doors
www.andersenwindows.com Cached

Fig. 3 Results obtained from three famous search engines for the query "window"

results obtained for the query "window" differently, because these search engines use different diversification techniques.

The purpose of this survey was to find the algorithms used for search result diversification; relate these algorithms with the type of queries; define a taxonomy based on: (1) the types of queries involved in search result diversification, and (2) mapping the diversification algorithms to the branches of taxonomy tree; discuss the evaluation metrics of the search result diversification methods; and lastly, present the future research directions in this area.

1.3 Proposed taxonomy

This survey presents a taxonomy of different diversification methods which handle different types of search queries. It has been observed during the study that there exist some basic diversification algorithms, and some variants of these algorithms are used to deal with different types of queries. The search queries have been classified into five main categories, and then, diversification methods have been associated with different classes of search queries. The proposed taxonomy based on the query classes and diversification algorithms has been presented in Fig. 4.

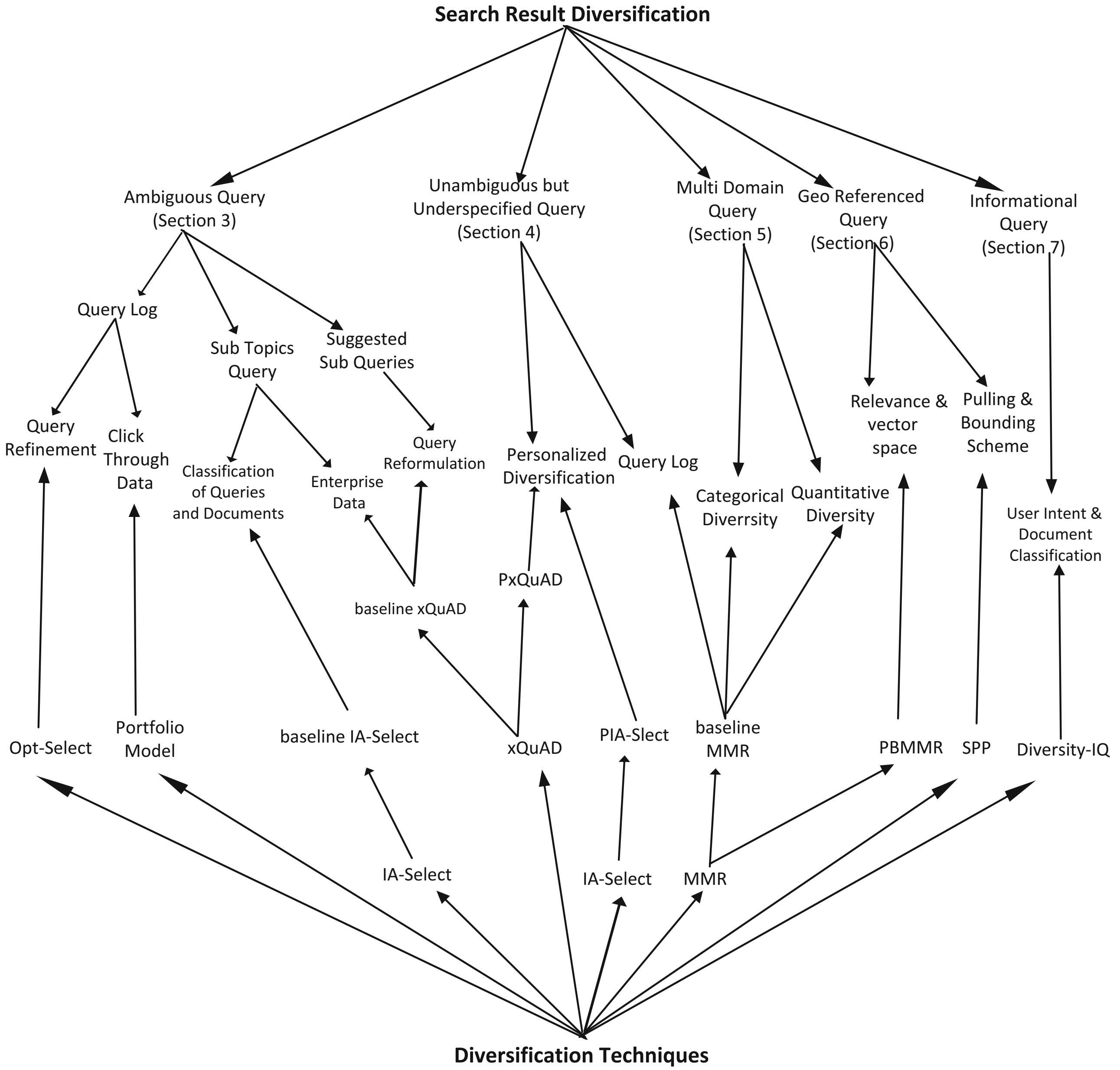

Search Result Diversification
Ambiguous Query (Section 3)
Unambiguous but Underspecified Query (Section 4)
Multi Domain Query (Section 5)
Geo Referenced Query (Section 6)
Informational Query (Section 7)
Query Log
Sub Topics Query
Suggested Sub Queries
Query Refinement
Click Through Data
Classification of Queries and Documents
Enterprise Data
Query Reformulation
Personalized Diversification
Query Log
Categorical Diverrsity
Quantitative Diversity
Relevance & vector space
Pulling & Bounding Scheme
User Intent & Document Classification
baseline xQuAD
PxQuAD
Opt-Select
Portfolio Model
baseline IA-Select
xQuAD
PIA-Slect
baseline MMR
PBMMR
SPP
Diversity-IQ
IA-Select
IA-Select
MMR
Diversification Techniques

Fig. 4 Taxonomy of search queries and mapping of diversification methods onto the query classes

The classification of search queries in this taxonomy is as follows:

- *Ambiguous query* An ambiguous query is the one which has more than one meaning. Literature reveals that three different techniques have been used to handle ambiguity in a query. First technique uses *a subquery* to discover different characteristics of the original query in the form of subquery. The second technique is *topic-based short query*, which uses query expansion methods to close the gap between brief expressions and retrieval goals. The third technique involves *query log* to handle the ambiguous query.
- *Unambiguous but underspecified queries* These types of queries are unambiguous in the logic that they do not have any contextual constraints. Although such queries are unambiguous in terms of topic, yet the user needs are not clear. There are two techniques to resolve such queries: the first one is *personalized diversification*, which is used to present the correct information to the correct person at the correct time; the other technique uses *log* to process the search result based on most frequently used pages.
- *Geo-referenced query* Geo-referenced query involves the user requests where the user finds relevant objects closer to a given location. There are two techniques to process such queries: first technique is based on the relevance and the vector space model that is used to retrieve most relevant results, and covers most neighborhoods. Another technique is pulling and bounding scheme where objects are contained in a finite bounded region.
- *Multi-domain query* Multi-domain query contains multiple linked concepts. These queries are resolved by two different techniques: one technique is *categorical diversity*, which relates two combinations based on the equality of the values of one or more categorical attribute of the tuples, while another technique is *quantitative diversity* that is used to measure diversity in terms of distance.
- *Informational query* The meaning of such queries is clear, but the query is justified by more than one result. Such queries are processed based on the relevant subtopics and the possibility of user's interest in these subtopics. Such techniques tend to produce an ordered set of documents so that an average user finds sufficient relevant documents.

1.4 Outline

The remainder of this article is organized as follows. Section 2 presents a general framework for diversification algorithms. Section 3 explains the class of ambiguous queries and associated methods for diversification. Section 4 presents the class of unambiguous but underspecified queries and maps the diversification methods on its different subclasses. The diversification techniques for multi-domain queries have been presented in Sect. 5, whereas Sect. 6 describes the diversification methods to handle the geo-referenced queries. The informational queries and relevant diversification techniques have been presented in Sect. 7. Lastly, Sect. 8 presents an overall discussion about search result diversification algorithms, diversity-aware evaluation measures, and dataset. It also presents the future directions about search result diversification.

2 Search result diversification framework

The main purpose of search result diversification was to find relevant and diverse result set for an input query. The literature survey divulges that the search results diversification framework is based on three components, namely relevance measure, diversity measure, and diversification objective. The first component produces the top most relevant results. The second component produces overall dissimilarity of the result set. The final component defines the ways with which both relevance and diversity merge into a single score that has to be maximized [1].

2.1 Relevance measure

The relevance measure is used to compute the similarity between a candidate document and the user input. This similarity is generally referred to as the relevance score, and an initial ranking of the results is based on this relevance score. There are many standard techniques which have been used to rank the items by their relevance, for example, vector space model to represent item and queries; language model [2]; KL divergence [3] which is used as relevance function.

2.2 Diversity measure

Diversity is closely related to the idea of similarity. Diversity is computed based on the similarity of documents within the result set, the more the documents in a result set are similar, the less diverse the result set is [4]. Furthermore, different notions of diversity have been investigated.

1. *Semantic distance* Sementic distance is used to measure the relevance between query and document. There are different techniques used in information retrieval for finding semantic distance [2] such as cosine similarity, Jaccard similarity [2], and Euclidean distance [5].

2. *Categorical distance* This type of distance is generally used in enterprise datasets, where the data objects are represented in a structured or semistructured formats, e.g., relational database or XML [6]. Categorical distance measures the similarity or distance between two attributes. There are different techniques for finding categorical distance, such as Manhattan distance and Supremum distance. For example, consider the relational database, in which order among attributes matters (e.g., for cars: Make -> Model -> color ->). This order expresses that certain attributes have priority to be diversified than other (e.g., first Make is diversified, then model). This shows that how result tuples can be seen as paths in a tree of values.

2.3 Diversification objective

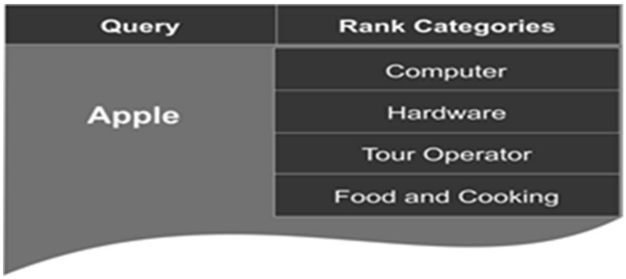
Search result diversification can be achieved by the relevance of query and documents and similarity between documents in the result set. The main objective of diversification was to find the optimal set of items, which is both relevant and diverse. This component formalizes the strategy to find a trade-off between the two measures in order to diversify a result set. The relevance and diversity of a search result can be combined by following different strategies.

1. *Max-sum diversification* This first objective was to compute the sums of relevance score of each document with the search query, it also computes the diversity of each document in the relevant result set. At the end, combine the relevance score and diversity as a weighted sum.
2. *Max-min diversification* The target of second objective is to increase the sum of those documents which have minimum relevance and maximum dissimilarity within the result set. Max-min diversification is important for those documents which have low relevance and diversity but may be important for the user.
3. *Average dissimilarity diversification* Here, the objective was to sum the original relevance for a result with the average dissimilarity of all documents in the result set. The main theme of average dissimilarity maximizes the sum over the whole set.
4. *Max-sum of max-score diversification* This function gives more importance to the relevance between query and documents. The max-sum of max-score produces a set of results that have the maximal relevance sum and then adds maximum diversity into final result set.
5. *Categorical diversification* This method is used to measure the relevance between the categories of documents and query. The result set is diversified if it covers all the categories of documents, and categories are weighted by their probability to occur.

3 Ambiguous query

Ambiguous queries have more than one meaning. It is generally supposed that many queries submitted to search engines are ambiguous [7]. For ambiguous queries, the search engine needs to ensure that the documents corresponding to different possible interpretations of the query should be presented to the user. In such a scenario, the search engine can present a set of results to the user that cover different aspects underlying the original user query. Consider, for example, the term "Apple" [8]. In Fig. 5, it is shown that the query "Apple" might refer to computer or to any hardware, or may refer to a famous tour operator in the USA. Without any further information, this query remains unclear and thus demands results from many different relevant perspectives. In order to process such queries effectively, the search engine should make a set of results possibly covering all (the majority of) the different understandings of the query. The problem of ambiguous queries has been addressed by using three different techniques.

1. *Query log* Web search engine (WSE) gathers complete information about submitted queries with the help of query log that are really valued for ambiguous query [9]. Such types of techniques are discussed in Sect. 3.1.
2. *Query subtopics* Ambiguous query should exploit the satisfaction of a user by covering a variety of subtopics in which a searcher could be interested. This method is used to find meaningful query subtopics [10]. This technique uses the previous information about subtopics of a query and statistical information about the user's intent on these subtopics. Such types of techniques are discussed in Sect. 3.2.
3. *Suggested subqueries* The submitted queries frequently transfer some ambiguity, this type of query can be broken into different subqueries. Actually, this technique is used to discover the different characteristics underlying the original query in the form of subqueries [11]. Such type of techniques is discussed in Sect. 3.3.



Query	Rank Categories
Apple	Computer
	Hardware
	Tour Operator
	Food and Cooking

Fig. 5 Result sets for query "Apple"

3.1 Query log

Query log of a *web search engine* keeps the amount of information about users' behavior, their requirements, and how users communicate with the search engines. Users click history may possibly be used for diversification. To this end, click entropy and query statistics are used to retrieve relevant result set from query log. Certainly, the usage of query log invites a discussion on privacy issues, however in general search engines use their own logs for the experimentation, whereas, the cases where the search engines expose their logs to the experimenters they ensure the anonymity of the user.

In Fig. 5, the user passes the query "apple." This ambiguous query goes to query log. Figure 6 presents a visualization of the idea of using query log, where the number of users clicking on each possible variant of the query "apple" have been presented, which helps rating different possibilities using the data available in the log. The figure shows that query log statistics reveal that seven users clicked on the Web site showing *apple Mac*, four users clicked on *big apple*, and one user clicked on apple *fruit*. With the help of a query log technique, the Web site showing *apple Mac* comes in the first place, followed by that of *big apple,* and at the end apple *fruit*.

3.1.1 Query refinement and Opt-Select diversification

Users interact with WSE through entering a few keywords, and these keywords are often ambiguous. WSEs also gather complete information about already submitted queries in the past along with extra information which are very useful for different tasks. Query log is used to return different search results to cover different interpretations of the query.

3.1.1.1 Query log-based documents utility This framework uses the following parameters [12, 13]

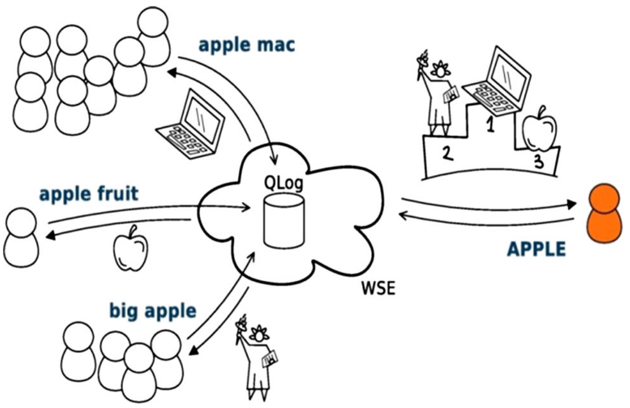


Fig. 6 Example of query log [12]

- D is the group of documents.
- Assumed query q.
- Rq is the set of documents that belong to a group.
- Sq is a set of possible query specialization.

The utility function, which helps identifying the novelty of a document, has been defined in Eq. (1).

$$U(d|Rq') = \sum_{d' \in Rq'} \frac{1 - \delta(d, d')}{\text{rank}(d', Rq')} \tag{1}$$

where Rq' is the list of results given by the search engine query q'

3.1.1.2 Opt-Select diversification algorithm The algorithm presented in Fig. 7 involves an original query q and result set Rq for a query q. Two probabilities are computed $P(d|q)$ and $P(d|q)$ and $P(q\prime|q)\forall q\prime \in Sq$, which are mixed by using parameter λ, where $\lambda \in [0, 1]$. The utility $U(d|Rq\prime)$ is used for documents. Here, the objective was to discover a set of documents $S \subseteq Rq$ with $|S| = k$ that maximizes the following expression in Eq. (2).

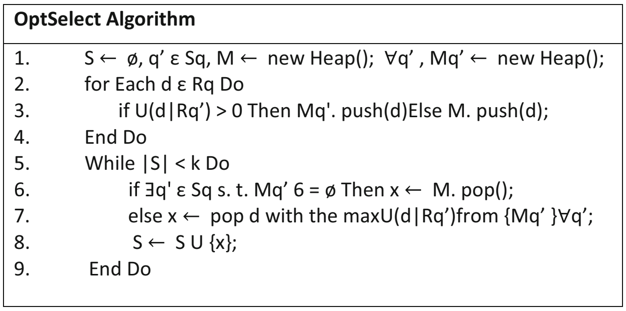
$$U(S|q) = \sum_{d \in S'} \sum_{q' \in S'} (1 - \lambda) P(d|q) + \lambda P(q'|q) U(d|R_{q'}) \tag{2}$$

In short, the Opt-Select [12] algorithm uses a query recommender system to obtain a set of queries for which Sq is built by including the most popular recommendation.

3.1.2 Click through rate and portfolio model

Click entropy is mostly used to identify queries that can be possibly benefited from search result diversification. Click entropy measures the variability of search results that a user clicks on (higher scores reproduce that user click on many results) [14, 15].

This framework uses the following parameters [9], S presents the result set, Qs is a set of searches, whereas S' is a portfolio and the size of S' is based on page layout. In this model, all searches are considered unique.



OptSelect Algorithm

```
S ← ∅, q' ε Sq, M ← new Heap(); ∀q' , Mq' ← new Heap();
for Each d ε Rq Do
    if U(d|Rq') > 0 Then Mq'. push(d)Else M. push(d);
End Do
While |S| < k Do
    if ∃q' ε Sq s. t. Mq' 6 = ∅ Then x ← M. pop();
    else x ← pop d with the maxU(d|Rq')from {Mq' }∀q';
    S ← S U {x};
End Do
```

Fig. 7 Opt-Select algorithm [12]

For example, the desired result for “Michael Jordan” is

1. Q1 search is “NBA” player
2. Q2 search is “University of Berkely Professor”

Let Z_u be a random variable indicating the relevance of document u to queries in Qs Now, let $Z = [Z_1, Z_2, \ldots, Z_n]$ be a vector of random variables indicating the relevance of documents in S This model denotes the correlation between two variables Z_i and Z_j by P_{ij} and from the covariance matrix of the variable associated with the result set S.

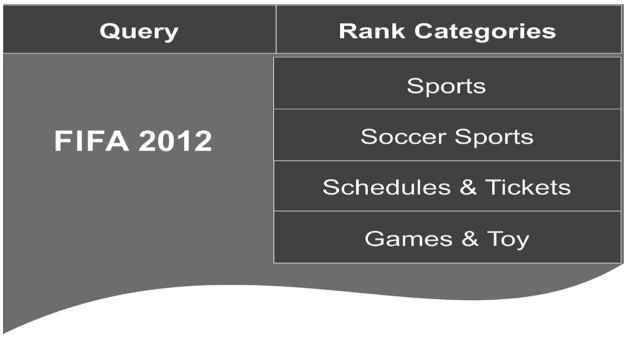
A portfolio is diversified if the most relevant and diversified results are displayed at the top. This algorithm is diversifying the search results presented by Google. The limitation of this model is that every topic should have a different page title.

3.2 Subtopics query

The objective of information retrieval was to provide related information to the users according to their searching keywords/topics. Based on the fact that users often issue very short queries, query expansion methods have been proposed to map the user’s queries to their retrieval objectives. By using topic-based queries, the same topic may be understood in different domains for various users. The simple idea after query expansion is to add extra subtopics to the topic-based queries so that the retrieval objective can be stated more specifically and accurately. Consider the query “FIFA 2012”; Fig. 8 shows that this query “FIFA 2012” is a topic-based query, and it might refer to sports, soccer sports, and schedules and tickets or to the games and toy.

3.2.1 Enterprise data and diversification

Enterprise data are used to mix the structured and unstructured data to determine query subtopics for search result diversification. A subtopic mined from structured data holds high-class terms, while the subtopic mined from unstructured data can well represent the document content that may cover a lot of noisy terms.



Query	Rank Categories
FIFA 2012	Sports
	Soccer Sports
	Schedules & Tickets
	Games & Toy

Fig. 8 Result set of the *FIFA 2012* query

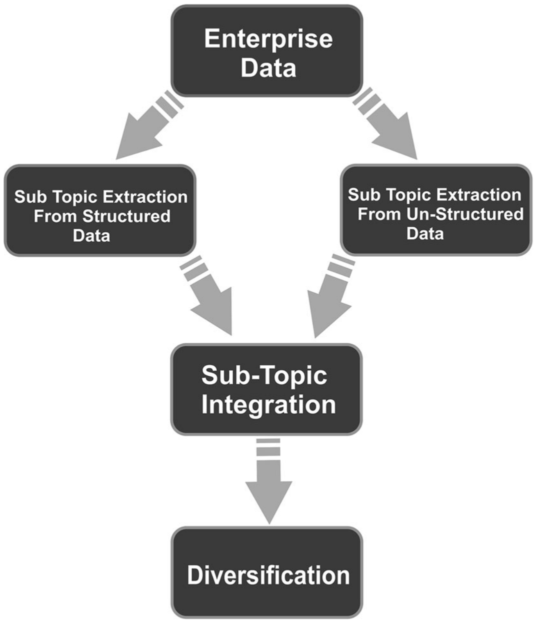

Enterprise Data
Sub Topic Extraction From Structured Data
Sub Topic Extraction From Un-Structured Data
Sub-Topic Integration
Diversification

Fig. 9 Workflow of subtopic generation in enterprise data

Thus, the enterprise data method is used to integrate the subtopic mined from structured data with the ones from unstructured data [10]. A workflow of subtopic generation has been presented in Fig. 9.

This framework retrieves relevant query subtopics and diversifies search result using the xQuAD diversification method.

3.2.1.1 Subtopic extraction from structured data This method uses the relational database model to extract subtopics from the structured data. The objective was to select K subtopics that cover dissimilar yet related information about the query. The score of the relevance of node s_i in the subtree rooted at s_i and the query q is given in Eq. (3).

$$\mathrm{rel}(si, q) = \frac{\sum s \in T_{si}^{\mathrm{sim}(s,q)}}{|T_{si}|} \tag{3}$$

Here, s_i is the ith node in the database structure. T_{si} is the subtree rooted at s_i, and q is the query, whereas $\mathrm{sim}(s, q)$, presented in Eq. (4), is the semantic similarity between s and the query q.

$$\mathrm{sim}(s, q) = \frac{\sum t \in s^{\mathrm{sim}(t,q)}}{|s|} \tag{4}$$

where t is the term in s. This framework iteratively selects K nodes having the highest scores in the subtopics.

3.2.1.2 Subtopic extraction from unstructured data This method uses probabilistic latent semantic analysis (PLSA) to mine subtopic information from unstructured data. In order to avoid the overlying information in different subtopics, the system allocates each term to the cluster which has the highest value in the PLSA result, as shown in Eq. (5).

$$S'(t) = \arg\max_{s'\in S'} \mathrm{score}(t, s') \tag{5}$$

where S' is the set of subtopics, t is the term, and $S'(t)$ is the subtopic that t is assigned to, while each PLSA cluster is a subtopic. These subtopics are extracted from the clusters of the documents.

3.2.1.3 Subtopic integration Subtopic integration gives k subtopics that are extracted from the database and documents. Figure 9 shows the working of subtopic integration. This method integrated K subtopics, where each subtopic contains M terms. Join each subtopic of databases with subtopic of document based on their semantic similarity.

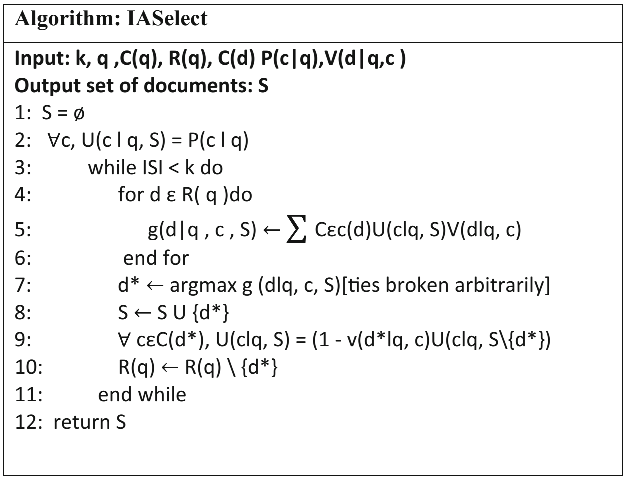
$$\mathrm{Si} = \arg\max_{s'\in S'} \mathrm{sim}(s, s'). \tag{6}$$

3.2.1.4 xQuAD diversification framework Relevant results are then diversified by xQuAD method which is discussed in Sect. 2.3.1. The score of each document in the result set is based on associations between document and query. Diversification method selects the documents in the result set that are similar to the query and subtopic.

3.2.2 Classification of queries and documents and IA-Select diversification

This technique is used to classify documents and queries into different categories. IA-Select, a search result diversification technique presented in Fig. 10, is used for the classification of documents and queries. This framework uses R that is a relevant result set based on the query q. Different parameters of this algorithm are as follows, q represents the initial query, c is the category to which q belongs to, and d is a document, whereas S is the diversified result set.

IA-Select computes the *conditional probability* U(c|q, S) between query q and category c. It selects the documents which have the *highest marginal utility,* which, in turn, is calculated by $g\ (d\prime|q,\ c,\ S)$. The drawback of this algorithm is that it is not optimal if a document belongs to more than one category.



Algorithm: IASelect

```
Input: k, q ,C(q), R(q), C(d) P(c|q),V(d|q,c )
Output set of documents: S
S = ø
∀c, U(c l q, S) = P(c l q)
     while lSl < k do
          for d ε R( q )do
               g(d|q , c , S) ← ∑ Cεc(d)U(clq, S)V(dlq, c)
          end for
          d* ← argmax g (dlq, c, S)[ties broken arbitrarily]
          S ← S U {d*}
          ∀ cεC(d*), U(clq, S) = (1 - v(d*lq, c)U(clq, S\{d*})
          R(q) ← R(q) \ {d*}
     end while
return S
```

Fig. 10 IA-Select algorithm [49]

3.3 Suggested subqueries

Web search engines normally offer suggested subqueries of the original query and subqueries helping the users to improve their original queries. This technique relies on finding different features which are essential for original query in the shape of subquery.

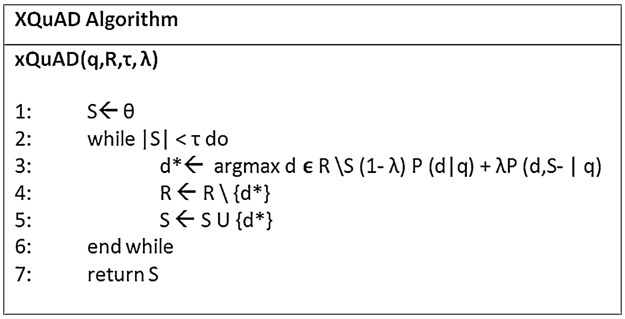
The literature shows that the problem of subquery generation has been addressed using three different techniques.

1. *Query expansion techniques* Query expansion is the process of reformulating a query to improve search results. This technique is used to assess a user's input and expanding the search query to match additional documents [16].
2. *Document clusters* A web search engine often returns thousands of pages in response to a broad query, making it difficult for users to browse or to identify relevant information. Clustering methods can be used to automatically group the retrieved documents into a list of meaningful categories [10].
3. *Query log* web search engines stored log information about users. This log information often serves to present different results of ambiguous queries. Query log is more general web search method.

3.3.1 Query reformulation and xQuAD diversification

This method is used to reformulate the original query, since the original query is not clear in its meaning. Query reformulation is used to cover different aspects of the original query [17]. Subquery generation method helps to complete the query reformulation framework. WSEs use query reformulation technique to determine different query aspects.

In the framework presented in [18], subqueries play a fundamental role. The algorithm is presented in Fig. 11 and is a probabilistic method of diversification.where q is ambiguous query; R is the initial ranking; τ represents the number of documents to be selected; S is the subset of ranking; λ controls the trade-off; $P(d|q)$ Given q, the probability with which document d is detected, and



XQuAD Algorithm

```
xQuAD(q,R,τ,λ)

1:   S ← θ
2:   while |S| < τ do
3:       d* ← argmax d ∈ R \S (1- λ) P (d|q) + λP (d,S- | q)
4:       R ← R \ {d*}
5:       S ← S U {d*}
6:   end while
7:   return S
```

Fig. 11 xQuAD diversification framework [18]

$P(d,\bar{S}|q)$ represents the probability of document d, but not the selected documents is observed, given the query q.

For each unselected document, it calculates the probability and chooses the document with the highest probability using the following expression presented in line 3 of Fig. 11. $(1-\lambda)P(d|q) + \lambda P(d,\bar{S}|q)$. where $(1-\lambda)P(d|q)$ is relevance, and $\lambda P(d,\bar{S}|q)$ is diversity. The algorithm adds them to subset list S, removes from original list R, and performs the same calculations until τ documents are collected.

This framework, which calculates $P(d,\bar{S}|q)$ and initial query, has multiple aspects presented in Eqs. (7) and (8):

$$P(d,\bar{S}|q_i) = \sum_{qi\in Q} P(q_i|q)P(d,\bar{S}|q_i) \tag{7}$$

where $P(q_i|q)$ reflects the importance of subquery q_i, and d is not dependent on the already selected documents.

$$P(d,\bar{S}|q_i) = P(d|q_i)P(\bar{S}|q_i) \tag{8}$$

where $P(d|q_i)$ represents coverage, and $P(\bar{S}|q_i)$ reflects the novelty of document d. Here, novelty is calculated by the probability of q_i not being satisfied by already selected documents (no need to compare document d to each of the selected documents).

Documents in S are independent from each other given the subquery q_i as shown in Eq. (9).

$$P(\bar{S}|q_i) = P\left(\overline{d_1,\ldots,d_{n-1}}|q_1\right) = \prod_{dj\in S}(1-P(d_j|q_i)) \tag{9}$$

Equation (10), which sums up the overall expression of this framework for document relevance and document diversity, is as follows:

$$\lambda \sum\nolimits_{q_i\in Q}\left[P(q_i|q)P(d|q_i)\prod\nolimits_{dj\in S}(1-P(d_j|q_i))\right] \tag{10}$$

Here, $(1-\lambda)P(d|q)$ reflects document relevance; $\lambda\sum_{q_i\in Q}[P(q_i|q)P(d|q_i)\prod_{dj\in S}(1-P(d_j|q_i))]$ represents the diversity of the document d; similarly, $P(q_i|q)$ reflects the importance of subquery; $P(d|q_i)$ addresses the coverage of the document, and $\prod_{d_j\in S}(1-P(d_j|q_i))$ represents the novelty of the document; Q is the subquery generation mechanism. The effectiveness of this algorithm can be made better by assessing the relative significance of each recognized subquery.

3.4 Discussion

An ambiguous query is the one which has more than one meaning. For example, "jaguar" can mean both an animal or a car. There are three different methods to deal with ambiguous query. First one is based on *suggested subqueries*; this type of query can be broken into different subqueries to discover the different characteristics of the original query in the form of subqueries. The second involves *query subtopics;* this method is used to find meaningful query subtopics and closes the gap between the query and relevant result set. Third one is based on *query log* which stores browsing information about users. Click entropy and query statistics are used to retrieve relevant result set from query log. Log information often serves to present different results of ambiguous queries. Three diversification algorithms xQuAD, IA-Select, and Opt-Select are used for this purpose.

The algorithm xQuAD involves subquery generation based on relevant documents and thus relates the relevant documents with appropriate subquery. This improves the processing of the results by avoiding matching documents to each other. The experiments show that xQuAD method generates effective subqueries.

IA-Select is specially designed for subtopics technique where documents and queries are classified according to these subtopics. The subtopics are generated by using query expansion techniques. The experimental results presented in the relevant literature reflect that this method does not perform very well for a document which belongs to more than one category.

Opt-Select is specially designed for manipulating the information extracted from query log. Query log is used to detect the submission of ambiguous queries in the past and is used to cover the possible interpretations of the query.

The research related to the ambiguous queries can be enhanced in the following directions: The ambiguous queries are generally processed using the query log, and the methods used to resolve such queries collect the statistics about most frequently accessed documents related to a query. The results of such methods can be improved by extracting more useful statistics about the documents, or by incorporating probabilistic measures over the gathered statistics.

The ambiguous queries are processed by using subtopic generation. For such queries, it is required to identify the relative importance of each subtopic and then involve the most important subtopics in the process of producing a diversified result set.

Subquery generation is another method to handle ambiguous queries, identifying meaningful subqueries and estimating their relative importance are challenging problems

4 Unambiguous but underspecified query

The sense of these queries is unambiguous and clear, and there is only one way to read or understand these queries. However, it is not clearly specified what the user wants to know about the entity. For example, consider the query “Madonna” presented in Fig. 12; the meaning of the query is clear but what the user wants to know about Madonna is not clear, does he need to look out the music videos, find lyrics of any song, purchase the songs at the iTunes store, or read news. In short, user’s interest is not specified. For such queries, the search engine needs to focus on determining the user’s interest behind the underspecified query and make a list of results that cover these dissimilar intents accordingly.

It has been observed that the problem of unambiguous but underspecified queries has been addressed using two different techniques:

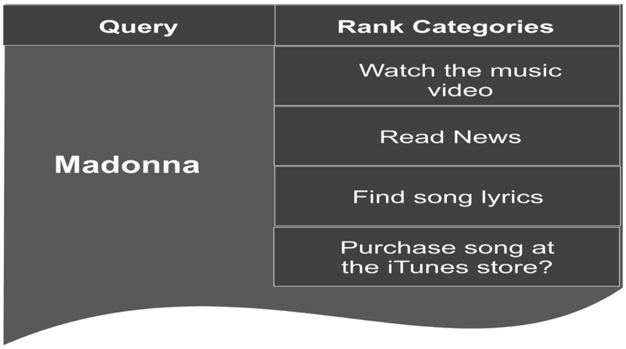
1. *Personalized diversification* This procedure constitutes two steps: first, this scheme gathers personal information from user profiles, and second, the diversification must be applied on the relevant result set [19–21]. Such type of technique is discussed in Sect. 4.1.
2. *Query log* In query log scheme, the system automatically suggests a set of queries, based on the original query; the proposed suggestion represents a different possible interpretation [22, 23]. Such kind of procedure is discussed in Sect. 4.2.

4.1 Personalized diversification

Personalization is the process of presenting the right information to the right user at the right moment. This method essentially gathers personal information, evaluates it, and then stores it in the user’s profile. For creating profiles, users select the categories of topics in which they are interested, and the search engine uses this information during the process of retrieval [24, 25].

There are two groups of user profiles

User’s preferences (e.g., search engines preferred, types of documents)
User’s interests (e.g., sports, photography).



Query	Rank Categories
Madonna	Watch the music video
	Read News
	Find song lyrics
	Purchase song at the iTunes store?

Fig. 12 Result set of the *Madonna* query

User profile A user profile is constructed from web pages browsed by the user. However, this technique focuses on using the user’s search history.

Profile based on user’s preferences User profile, based on the user’s preference, runs as a background process on the user’s machine. The application can retrieve results immediately after a query has been submitted. In this case, the profile is supplied to an agent that can automatically gather information on behalf of the user.

Profiles based on user’s interests This technique is based on users’ interest rather than users’ preferences. Such profiles are based on user’s browsing history. This system implicitly creates profile using browsing histories rather than explicitly created a profile [24].

Personalized web search model In this model, [26] requested query maps to user’s interests which are written in the user’s profile. The order of the personalized result set is the last step of the personalized web search model. Consider the query “Queen” in Fig. 13. The result set of the query is based upon user profile and personalized ordering.

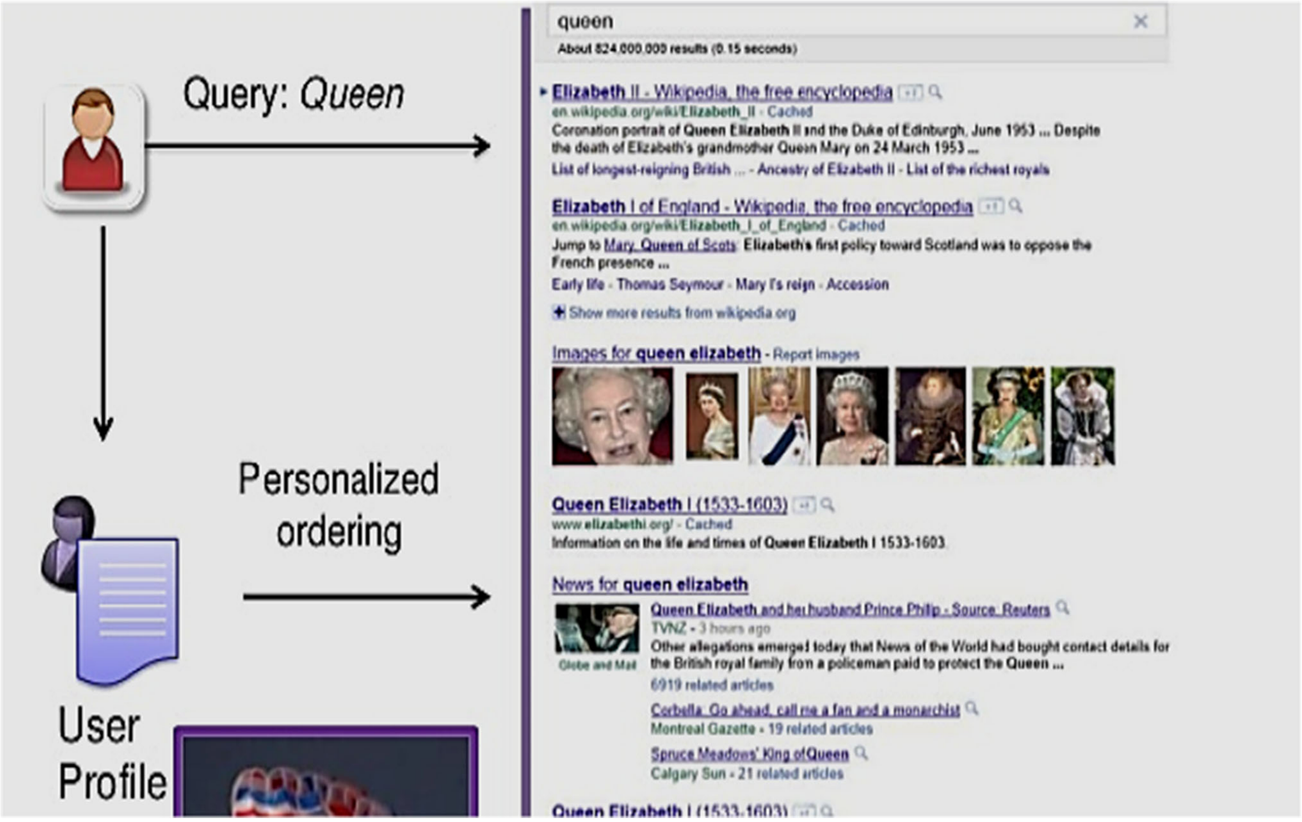
Diversified web search model Figure 14 shows the result set of query “Queen” by using diversification model. One positive aspect of this model is that all links are relevant and almost different from each other.

4.1.1 *Diversify personalization framework*

Probabilistic model The diversity personalization framework is based on a probabilistic model which involves the following expressions:

$p(c|q)$: Relation between category and the query (e.g., popularity of certain aspect in a query)
$p(q|d), p(d|q)$: Relation between document and query (e.g., ranking score of document)
$p(c|q), p(d|c)$: Relation between document and category (e.g., document classification)

IA-Select This framework is based on Eq. (11) which is as follows [19]



queen

About 824,000,000 results (0.15 seconds)

Elizabeth II - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Elizabeth_II - Cached
Coronation portrait of Queen Elizabeth II and the Duke of Edinburgh, June 1953 ... Despite the death of Elizabeth's grandmother Queen Mary on 24 March 1953 ...
List of longest-reigning British ... - Ancestry of Elizabeth II - List of the richest royals

Elizabeth I of England - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Elizabeth_I_of_England - Cached
Jump to Mary, Queen of Scots: Elizabeth's first policy toward Scotland was to oppose the French presence ...
Early life - Thomas Seymour - Mary I's reign - Accession

Show more results from wikipedia.org

Images for queen elizabeth - Report images

Queen Elizabeth I (1533-1603)
www.elizabethi.org/ - Cached
Information on the life and times of Queen Elizabeth I 1533-1603

News for queen elizabeth

Queen Elizabeth and her husband Prince Philip - Source: Reuters
TVNZ - 3 hours ago
Other allegations emerged today that News of the World had bought contact details for the British royal family from a policeman paid to protect the Queen ...
6919 related articles

Globe and Mail

Corbella: Go ahead, call me a fan and a monarchist
Montreal Gazette - 19 related articles

Spruce Meadows' King ofQueen
Calgary Sun - 21 related articles

Queen Elizabeth I (1533-1603)

Fig. 13 Working of personalized web search model

$$fs(d) = \sum_{c} p(q|d)p(c|d)p(c|q) \prod_{d' \in S} (1 - p(q|d')p(c|d')) \tag{11}$$

where

$p(q|d)p(c|d) =$ Document relevance

$p(c|q) \prod_{d/\in S} (1 - p(q|d')p(c|d')) =$ Novelty

Personalized IA-Select Adding a user component results into Eq. (11) results into [19]

$$\begin{aligned} fs(d) = & \sum_{c} p(q|d,u)p(c|d,u)p(c|q,u) \\ & \times \prod_{d' \in S} (1 - p(q|d',u)p(c|d',u)) \end{aligned} \tag{12}$$

xQuAD Equation (13) shows the expression for xQuAD algorithm [19]

$$\begin{aligned} fs(d) = & (1 - \lambda)p(d|q) \\ & + \lambda \sum p(c|q)p(d|c) \prod_{d' \in S} (1 - p(d'|c)) \end{aligned} \tag{13}$$

where $p(d|q)$ represents the relevance of document to the query; $\sum_{c} p(c|q)p(d|c)$ reflects the relevance of the document to the topic; $\prod_{d' \in S} (1 - p(d'|c))$ is the novelty; and (λ) provides the adjustment factor for the degree of diversification.

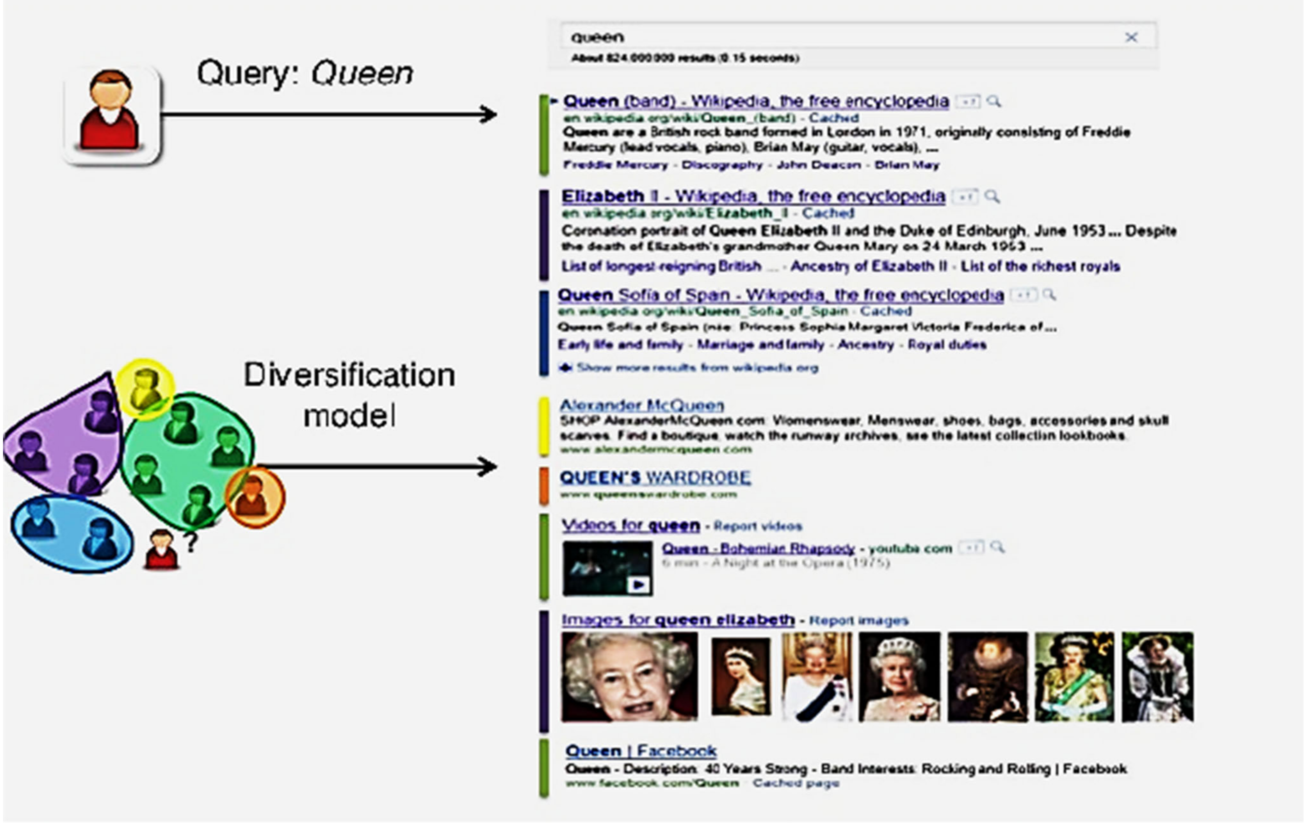
Personalized xQuAD Adding a user component results into Eq. (14) [19]

$$\begin{aligned} fs(d,u) = & (1 - \lambda)p(d|q,u) \\ & + \lambda \sum_{c} p(c|q,u)p(d|c,u) \prod_{d' \in S} (1 - p(d'|c,u)) \end{aligned} \tag{14}$$

4.2 Query log

Query log collects information from search history, user profiles, or user click history. By using personalization, following issues can rise:

1. It may be difficult or impossible to collect information or data from the user's to effectively build their profile
2. Gathering such data usually violates user privacy



queen

About 824,000,000 results (0.15 seconds)

Queen (band) - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Queen_(band) - Cached
Queen are a British rock band formed in London in 1971, originally consisting of Freddie Mercury (lead vocals, piano), Brian May (guitar, vocals), ...
Freddie Mercury - Discography - John Deacon - Brian May

Elizabeth II - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Elizabeth_II - Cached
Coronation portrait of Queen Elizabeth II and the Duke of Edinburgh, June 1953 ... Despite the death of Elizabeth's grandmother Queen Mary on 24 March 1953 ...
List of longest-reigning British ... - Ancestry of Elizabeth II - List of the richest royals

Queen Sofia of Span - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Queen_Sofia_of_Spain - Cached
Queen Sofia of Spain (née: Princess Sophia Margaret Victoria Frederica of ...
Early life and family - Marriage and family - Ancestry - Royal duties
Show more results from wikipedia.org

Alexander McQueen
SHOP AlexanderMcQueen.com: Womenswear, Menswear, shoes, bags, accessories and skull scarves. Find a boutique, watch the runway archives, see the latest collection lookbooks.
www.alexandermcqueen.com

QUEEN'S WARDROBE
www.queenswardrobe.com

Videos for queen - Report videos
Queen - Bohemian Rhapsody - youtube.com
6 min - A Night at the Opera (1975)

Images for queen elizabeth - Report images

Queen | Facebook
Queen - Description: 40 Years Strong - Band Interests: Rocking and Rolling | Facebook
www.facebook.com/Queen - Cached page

Fig. 14 Result set of diversified web search model

Query logs propose different queries, every suggested idea gives a different possible interpretation [22] of the user query.

An algorithm based on time succession This framework [22] uses automatic query suggestion that is based on historical query logs. In this framework, a graph is built, and each vertex of the graph represents a unique query from the logs.

4.2.1 *Maximal marginal relevance (MMR) diversification framework*

MMR method is used to diversify the result set. By using the original formulation of MMR, the framework proposes to adopt this method to the problem of diversified query suggestion. Thus, the adopted formula shown in Eq. (15) is as follows:

$$\text{qSuggMMR} = \arg\max_{qi \in R \setminus S}\left[\lambda \text{sim}_1(q_i, q) - (1-\lambda)\max_{qj \in S}\text{sim}_2(q_i, q_j)\right] \quad (15)$$

where q is the underspecified query, R is the set of candidate query suggestions, and q_i represents the candidates who are selected from an absolute set of query suggestions S The query suggestion method described in Sect. 3.1 is used to calculate R, that is, the set of candidates.

4.3 Discussion

Underspecified queries are unambiguous in the sense that the meaning of this query is clear, but it is difficult to figure out what details does the user require about the input query. Consider the query "friendships poem". Here, the meaning of the query is unambiguous but still it is not clear what the user wants to know about friendship poem. The unambiguous but underspecified query is processed using two techniques, namely personalized diversification and query logs. Personalized diversification has two steps: first, this method gathers personal information from user profile and then maximizes the probability of showing an interpretation relevant to the user. User profile based on user's preferences (search engines preferred types of documents based on browsing history) and a user's interest (categories of topics in which user is interested). If a user profile is perfectly defined, then personalization diversification approach gives relevant and diversified result set. If a profile has errors, then personalized diversity is preferred over full personalization or diversification.

The query log approach automatically suggests a set of queries, based on the original query. Query log collects information from search history, user profiles, or user click history. In personalization method, it may be difficult or impossible to collect information or data from the user's to effectively build their profile.

PxQuAD and PIA-Select diversification algorithms are used for the personalized diversification method. An MMR diversification algorithm is used for query log. PxQuAD performs better for subquery generation using user profile, and PIA-Select does well for subtopics. Another method to process such queries is MMR diversification algorithm. Unlike the previous approaches, this algorithm does not count on the user profiles. It uses the concept of text-based similarity such as the vector space model and generates candidate queries from the query log.

In future, the methods that deal with unambiguous but underspecified queries can be enhanced in the following ways. Diversity and personalization can be joined in different ways, which provides a wide room of future research. For instance, there is a need to work on the exaggerated use of user's search history for personalizing resultant diversification.

5 Multi-domain query

Multi-domain search attempts to answer the queries that contain multiple linked concepts [27, 28] and spans across multiple things, i.e., these types of queries give answer by linking knowledge from more than one domain [29].

Example Search for upcoming concerts close to an attractive location (*like a beach, lake, mountain, natural park, and so on*), considering also the availability of good, close-by *hotels*…

In the above query, the search needs to be expanded to get information about available *restaurants* near the candidate concert locations, *news* associated with the events, and possible options to combine further *events* scheduled on the same days [29].

For example, consider the query (*concert, restaurant, and news close to an attractive location*). In Fig. 15, different result sets are presented for each term of the query. By using the multi-domain technique, all these different result sets are combined into a single result set; thus, the result of a multi-domain query comes from multiple pages.

The literature reveals that the problem of multi-domain queries has been addressed by using two different techniques.

1. *Categorical diversity* In this technique, the result set is selected on the equality of the values by using the technique of relational database [30], and this technique is discussed in Sect. 5.1.
2. *Quantitative diversity* In this technique, multi-domain query that need two or more combinations, the diversity of these combinations is defined by their distance, and the detailed procedure is discussed in Sect. 5.1.

Multi-domain queries are represented as a set of relations. All items of the result set are a group of different objects that satisfy the join and selection conditions, and the result set is ranked according to the scoring function

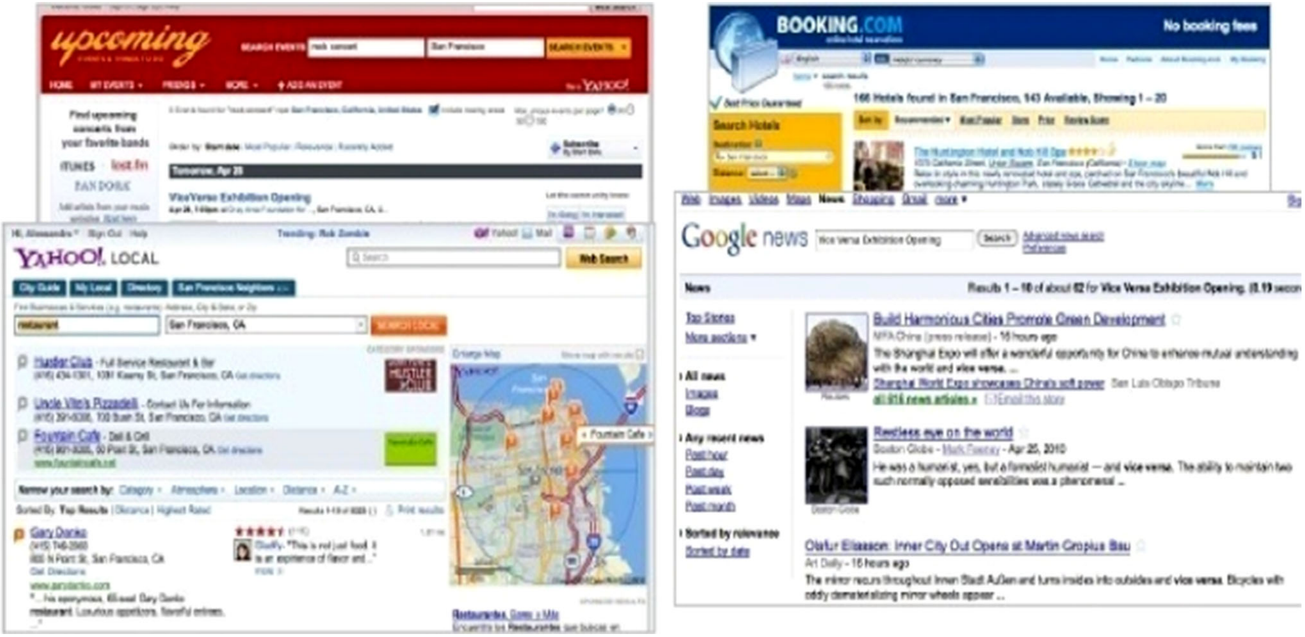


Fig. 15 Separate result sets of each term for multi-domain query

[31]. Due to the combinatorial nature of multi-domain search, the number of combinations in the result set is normally very high.

There are two criterions for comparing combinations. The details of these two criterions are discussed in Sect. 5.1.

5.1 Categorical and quantitative diversity

This method is used only when values declare only equality test. In categorical diversity, two combinations are related, based on the equality of the values of one or more categorical attribute of the tuples that establish them. Categorical diversity can be based on the key attribute [32].

However, the quantitative diversity of two combinations is defined as their distance, expressed by some metric function. This technique is helpful if the user wants to search result set near to his location. This method can improve the quality of multi-domain result set.

5.1.1 Multi-domain diversification

Consider a set of relations $R_1, R_2, \ldots, R_n$, where each R_i denotes the result set returned. Each tuple $t_i \in R_i$ has schema $R_i\left(A_i^1 : D_i^1, \ldots, A_i^{m_i} : D_i^{m_i}\right)$, where $A_i^{m_i}$ is an attribute of relation R_i and $D_i^{m_i}$ is the associated domain. This framework distinguishes the domains D_i^k into categorical diversification when values admit only equality test and quantitative diversification when values can be organized into vector embedded in a metric space. A multi-domain query over the search services is defined as a join query $q = R_1 \propto \cdots \propto R_n$ over the relations $R_1, R_2, \ldots, R_n$, where they can be joined using any arbitrary join predicate.

5.1.2 Relevance

The goal of multi-domain search was to select one or more combinations from the result set. User-defined relevance score function is $S(\tau, q)$ where q is the query, and τ is join condition. Scoring function is normalized in the [0, 1] range, where 1 indicates the highest relevance, when the result set $\mathcal{R}$ sorted, e.g., in descending orders of relevance.

5.1.3 Example

Given the relations

$\boldsymbol{Hotel}(\boldsymbol{HName}, \boldsymbol{HLoc}, \boldsymbol{HRating}, \boldsymbol{HPrice})$

$\boldsymbol{Restaurant}(\boldsymbol{RName}, \boldsymbol{RLoc}, \boldsymbol{RRating}, \boldsymbol{RPrice})$

$\boldsymbol{Museum}(\boldsymbol{MName}, \boldsymbol{MLoc}, \boldsymbol{MRating}, \boldsymbol{MPrice})$

Consider a function $\boldsymbol{city}()$ which takes geographical coordinates as input and returns the name of the corresponding city, and a multi-domain query q:

$$\boldsymbol{q} = \boldsymbol{select} \times \boldsymbol{fromHotel}, \boldsymbol{Restaurant}, \boldsymbol{Museum},$$

where

$$\begin{aligned} \boldsymbol{city}(\boldsymbol{Hloc}) &= \boldsymbol{Milan} \wedge \boldsymbol{city}(\boldsymbol{RLoc}) \\ &= \boldsymbol{city}(\boldsymbol{HLoc}) \wedge \boldsymbol{city}(\boldsymbol{MLoc}) = \boldsymbol{city}(\boldsymbol{HLoc}). \end{aligned}$$

The overall price of the combination $\boldsymbol{S}(\boldsymbol{\tau}, \boldsymbol{q}) = \boldsymbol{sum}(\boldsymbol{HPrice}[\boldsymbol{t_h}], \boldsymbol{RPrice}[\boldsymbol{t_r}], \boldsymbol{MPrice}[\boldsymbol{t_m}])$. This example could be used to rank hotel, restaurant, and museum triples. Note that S is a simple linear function based solely on a subset of the attribute values of the tuples which construct a triple.

5.1.4 Diversity

As stated previously in Sect. 5.1, there are two different criterions to express the similarity of combinations

1. Categorical diversity
2. Quantitative diversity

In both cases, for each pair of combinations τ_u and τ_v, it is possible to define a diversity measure $\delta : \mathcal{R} \times \mathcal{R} \longrightarrow [0, 1]$, normalized in the [0, 1] interval, where 0 indicates maximum similarity, and $\mathcal{R}$ is result set.

5.1.5 Computing relevant and diverse combinations

1. $N = |\mathcal{R}|$ denotes the number of combinations in the result set
2. $\mathcal{R}_K \subseteq \mathcal{R}$ is the subset of combinations that are presented to the user, where $K = |\mathcal{R}_K|$

This framework is interested in identifying a subset $\mathcal{R}_K$ which is both relevant and diverse. Fixing the relevance score $S(., q)$, the dissimilarity function $\delta(., .)$ and a given integer K result into Eq. (16).

$$\mathcal{R}_K^* = \overset{\arg\max}{\mathcal{R}_K \subseteq \mathcal{R}, |\mathcal{R}_{K|=K}} F(\mathcal{R}_K, S(., q), \delta(., .) \tag{16}$$

where $F(.)$ is an objective function, which contains relevance and diversity.

5.1.6 MMR diversification

Another objective function, closely related to the above-mentioned functions, is MMR [33]. MMR implicitly maximizes a hybrid objective function, whereby the relevance scores are added together, while the minimum distance between pairs of objects is controlled.

5.2 Discussion

Multi-domain search is used to answer the queries that have more than one entity, such as "Find a hotel in Milan close to a concert venue, a museum and a good restaurant". Multi-domain can be represented as a join query over a set of relations. Multi-domain search result sets have normally very high combinations and also have strongly relevant objects repeated with many other concepts. Thus, the user should require scrolling down the list of results to see different alternatives. The literature shows that two techniques have been used to resolve such queries, namely *categorical* and *quantitative* diversity.

Categorical method is based on measuring the equality of the value by using the technique of relational database. The data objects in this approach are represented in a structured or semistructured formats, e.g., relational database or XML. The categorical method measures the similarity between two attributes, whereas the quantitative method is used to measure the distance and retrieve the objects near to the user's location. This method is used to improve the quality of the result set. These two methods help computing the relevance of the data objects to the multi-domain query.

In terms of the diversification algorithms, MMR algorithm is used in case of *categorical* diversity relied on *relevance*, and in the case of *quantitative* diversity, it utilizes the *distance* between pairs of objects. The quality of the result set can be improved by keeping balance between relevance and diversity.

In future, there is a need to develop new diversification approaches for multi-domain query, which may involve the relative importance of a data source and may also involve semantics to get more reasonably diversified result set.

6 Geo-referenced query

Geo-referenced data are becoming increasingly prominent on the existing web, particularly after the provision of several location-based services. Geo-referenced data focuss commonly on finding relevant objects close to a given location. For geo-referenced data, diversification is useful where objects can be defined using the following properties:

1. A score
2. A two- or three-dimensional feature vector

Consider a user who moves to a new city and he wants to take an overview of real estate. The required result will be based on the following criteria [34, 35]

1. Relevance (price, square meter, etc.).
2. Coverage of neighborhoods

Example 1 Consider a *real estate query*: a sample search in a commercial service for flats in London between £200,000 and £300,000 returned 60,000 + results; if the user wants to browse just a few dozen of them in diverse neighborhoods, the system needs to access and present a number of objects proportional to the user's wishes, scattered throughout the London region, without accessing all the 60,000 + relevant apartments [36].

Example 2 Consider a query: where *the user is looking for a restaurant in Milan*. Figure 16 shows the top 15 diversified result set of the query over the region. It is clear in the figure that without using location-based service, it may be possible that the top 15 results point the same location. The result of the query by using diversification over the region points all locations near to the user.

The problem of geo-referenced queries has been addressed using two different techniques.

Pulling and bounding scheme In this method, the query selects a finite set of relevant objects, and vector space is

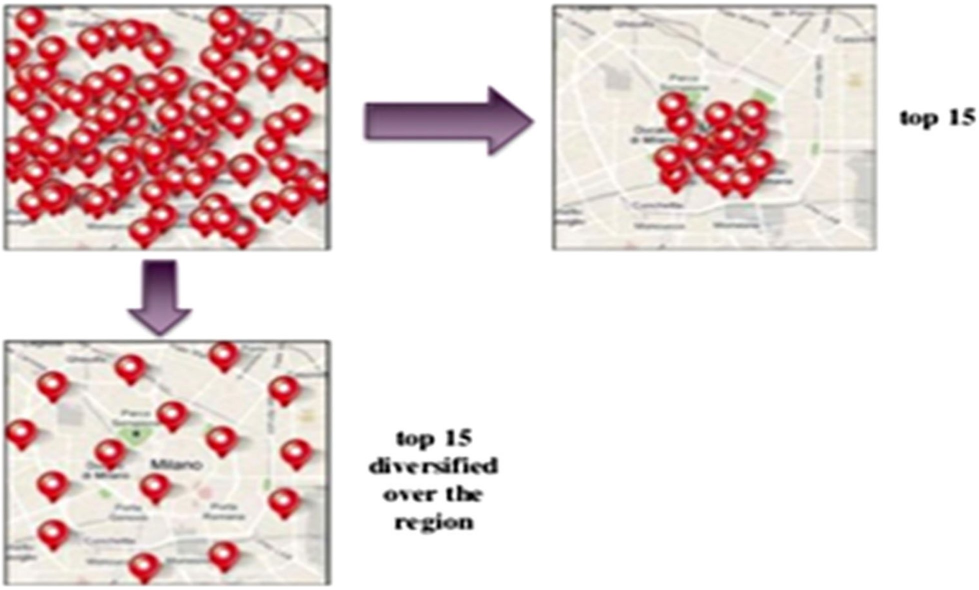


Fig. 16 Diversified result set over the region [36]

used to represent objects on the basis of distance [36]. Such type of technique is discussed in Sect. 6.1.

Pulling and bounding scheme In this technique, the objects are contained in a finite boundary region, and objects are fetched using score-based access and distance-based access [36]. This procedure is discussed in Sect. 6.2.

6.1 Probing location and relevance

Consider a query q selecting a finite set $\mathcal{O}$ of N object. The relevance of an object $o \in \mathcal{O}$ to q represented by a score $S_q(o) \in \mathbb{R}$. The vector space representation of an object $o \in \mathcal{O}$ to q is represented by a distance $X(o) \in \mathbb{R}^d[1]$. In this case, the diversification problem is solved using Eq. 17.

$$\mathcal{O}_K^* = \underset{\mathcal{O}_K \subseteq \mathcal{O}, |\mathcal{O}_K| = K}{\arg\max} F(\mathcal{O}_K; S_q(.), \delta(.,.)) \qquad (17)$$

where $\mathcal{O}_K^*$ represents the best diversified set of K objects; $\mathcal{O}$ is the set of objects; the objective function is represented by $F(\mathcal{O}_K; S_q(.), \delta(.,.))S_q(.)$ represents the relevance to query (as score); and $\delta(.,.)$ represents the diversity (as distance).

6.1.1 *MMR diversification framework*

MMR is the most popular algorithm with good quality of result (i.e., value of the objective function). This algorithm [36] finds K objects that are both relevant and diverse. At each step, pick the object with largest diversity-weighted score. The total numbers of steps are K.

$$\sigma(o; \mathcal{O}_K) = (1-\lambda)s_q(o) + \lambda^{\min}_{o\prime \in \mathcal{O}_K} Y\delta(o, o\prime) \qquad (18)$$

Equation (18) $\sigma(o; \mathcal{O}_K)$ represents the weighted score of diversity; $s_q(o)$ reflects the relevance; λ defines the trade-off between relevance and diversity; and ${}^{\min}_{o' \in \mathcal{O}_K} Y\delta(o, o')$ is the diversity. However, the corresponding objective function presented in Eq. (19) is as follows:

$$F(\mathcal{O}_K) = (1-\lambda) \sum_{O \in \mathcal{O}_K} s_q(o) + \lambda^{\min}_{o_u, o_v \in \mathcal{O}_K} \delta(o_u, o_v) \qquad (19)$$

The limitation of this algorithm is that all objects must be there from the beginning.

6.1.2 *Pull/bound maximum marginal relevance (PBMMR) diversification*

6.1.2.1 Framework This algorithm achieves the same quality of results as MMR. One of the key points of this framework is to reduce the number of accessing objects [36]. This algorithm uses k iterations, and each iteration makes the following two points as long as needed.

1. Pulling strategy:
    - Choose an access method (by score or distance)
    - If it chooses distance method, then select the probing location (i.e., from which point)
2. Bounding scheme:
    - This scheme computes an upper bound on the diversity-weighted score.

6.2 Pulling and bounding scheme

The objects are contained in a finite boundary region. Retrieving the objects is expensive, that is why objects are increasingly accessed, and the amount of accessed objects should be reduced [37, 38].

There are two categories of sorted access methods for fetching the objects.

6.2.1 *Score-based access*

The set O is accessed sequentially in decreasing order of relevance to the query, e.g., restaurant by score. In Fig. 17, the service *Restaurants* will be accessed by using a score-based access technique. A larger size of fork and spoon represents higher rating/score of a restaurant.

6.2.2 *Distance-based access*

The set O is accessed sequentially in increasing order [39] of distance from a given point. Figure 18 presents a diagrammatic description of accessing restaurants by using distance-based access technique.

6.2.3 *Space partitioning and probing (SPP) framework*

This method explores the region of space that grants the highest chances to retrieve the object with the best diversity-weighted score [36]. In each of the K iterations of the framework, it fixes the probing locations of the most

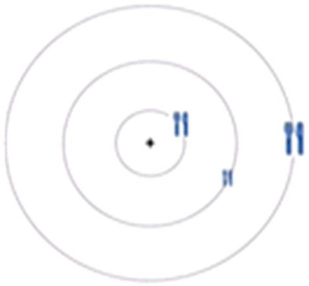


Fig. 17 Restaurant accessed by *score*

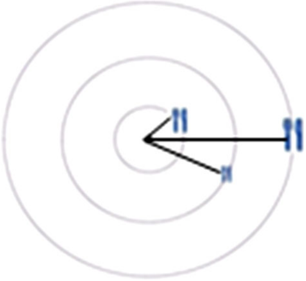


Fig. 18 Restaurant accessed by *distance*

promising points of the unexplored space. The vertices of the bounded Voronoi diagram [40] of the points selected at previous probing locations are points that lie within a bounding region U and are as far as possible from all the objects of the current selection $\mathcal{O}_1$

6.3 Discussion

Geo-referenced query involves the user requests where the user finds relevant objects closer to a given location, whereby user can create content attached to places. Geo-referenced data are very prominent on the web, after the beginning of location-based services. Geo-referenced query attaches content to places and is also found in domains such as *trip planning, news analysis, and real estate*. The basic purpose of geo-referenced data was to find relevant objects close to user location. Geo-referenced queries require the uniform coverage of a region. The diversification of geo-referenced queries is defined by relevant score and two- or three-dimensional vector.

The problem of geo-referenced queries has been addressed using two different techniques. First technique Relevance and vector space used PBMMR diversification algorithm. PBMMR uses MMR as a model for evaluating the quality of diversification. The goal of PBMMR was to achieve the same quality of results but to minimize the number of accessed objects. The pruning of data objects is conducted using the geometry of vector space for bounding scheme.

Pulling and bounded scheme is used in the SPP diversification algorithm. This method uses distance- or score-based data access, and objects are contained in a finite boundary region. It discovers the region of space that grants the maximum chances to retrieve the object with the highest diversity-weighted score. Thus, it reduces the number of accessed objects. The literature review reveals that SPP is very effective in reducing the number of objects being accessed.

In future, there is a need to work on tackling the possible presence of uncertainty in the data while applying diversification techniques for geo-referenced query [41].

7 Informational query

Informational queries are clear on the meaning and are properly specified, but the user supposes more than one result for her requirement. Such queries demand new and nonredundant information that involve different documents. User desires several results and uses these results for gathering information. For informational query, both novelty and redundancy are the most important [42]. For such queries, the user does not know in advance which document has exact information. Users certainly review a result set of informational query more in depth [43]. Consider, for example, the query shown in Fig. 19, "how to make cheesecake."

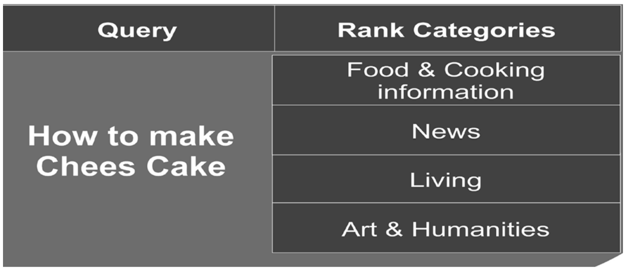
The meaning of the query is clear, but users often need more than one document to fulfill their information requirement; thus, the diversification method provides multiple relevant documents to help the users.

The problem of informational query has been addressed by using the following technique.

User intent and document classification In this technique, informational queries are resolved by finding relevant subtopics by using the possibility of user concern in all the subtopics. Document ranking is created with this method, after that average user finds sufficient related documents [44].

7.1 User intent and document classification

This method is used for identifying the relevant subtopics and document ranking, where documents are ranked by the probability of user interest in each of relevant subtopics. This method uses probability of information about query intent and relevance of documents with query subtopics. Based on the query intent probability, the system can classify which subtopics are significant for the users [44] (classification of document possibilities helps in the approximation that how likely a document is to satisfy a particular subtopic.)



Query	Rank Categories
How to make Chees Cake	Food & Cooking information
	News
	Living
	Art & Humanities

Fig. 19 Result set of "how to make cheesecake" query

7.1.1 *Relevant document requirements*

It is important for informational query to consider the number of relevant documents. For example, ten relevant documents most users visit. Users U usually need j documents connected to their subtopic [44] using the following expression.

$$\Pr(J=j|U), \quad \text{for } j>0.$$

7.1.1.1 User intent In this method, user issues search query T that has m subtopics $T_{1,}T_{2,}\ldots T_m$. User U is interested in subtopic T_i with probability $\Pr(T_i|U)$ [44].

7.1.1.2 Document categorization In this technique, document categorization is based on a probability distribution that document **d** is related to the topic **T**. For example, subtopics distribution **d** is relevant to $\mathbf{T_i}$ with probability $\mathbf{Pr(T_i|d)}$.

7.2 Diversification model of informational query

This method gives additional documents from a popular subtopic. To decide which documents are the best, probability distribution is the number of expected hits [45]. In this case, the query processing system should know prefect knowledge of user intent and document classification.

7.2.1 *Prefect knowledge of user intent*

Firstly, consider which subtopic T_i a user is interested in; on the contrary, the classifications of documents are probabilistic [46]. The number of documents that the user requires, denoted by j, must be considered in this technique, whereas the relevant documents are denoted by k. This method calculates the estimated amount of hits $E(R)$ for a set of n documents, as shown in Eq. (20).

$$E(R)=\sum_{j=1}^{n}\Pr(J=j|U)\sum_{k=1}^{n}\Pr(K_i=k|R)\min(j,k) \tag{20}$$

In the above equation, K_i is defined as the event that k documents in R belongs to T_i [44].

7.2.2 *Prefect document classification*

In this method, every document is categorized into one subtopic category, but the intent of the user is not known [47]. This method firstly considers the amount of documents selected from subtopic T_i as K_i and uses that for the m subtopics of T. This technique computes the number of expected hits of an average. Equation (21) effectively provides the relative importance of a subtopic/intent.

$$E(R)=\sum_{j=1}^{n}\sum_{i=1}^{m}\Pr(T_i|U)\Pr(J=j|U)\min(j,K_i) \tag{21}$$

where T_i represents the number of subtopics; U is the user; j is the desired number of documents by the user; and k is the number of relevant documents presented.

7.2.3 *Diversity-IQ diversification framework*

The two Eqs. (20) and (21) are combined to create probability distribution of above two sessions (perfect knowledge of user intent and perfect document classification). Equation (22) presents the combined expected number of hits.

$$\begin{aligned} E(R)=\sum_{j=1}^{n}\sum_{i=1}^{m}&\Pr(T_i|U)\Pr(J=j|U)\\ &\times\sum_{k=1}^{n}\Pr(K_i=k|R)\min(j,k) \end{aligned} \tag{22}$$

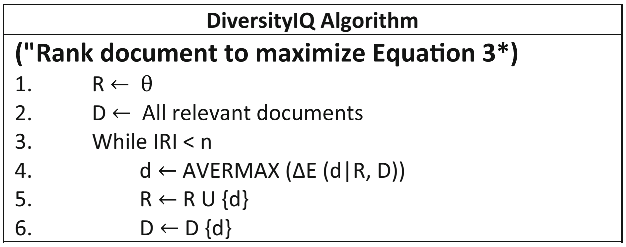
Diversity-IQ [44] algorithm, presented in Fig. 20, determines the set of documents R such that it maximizes the number of expected hits for an informational query.

The ΔE document computation is useful for different factors: firstly, for its subtopic scores; secondly, the interest of user in those subtopics; and thirdly, to compute the conditional probabilities to measure how various documents from every subtopic are relevant which are previously involved in R.

7.3 Discussion

The meaning of informational queries is clear, but the query is justified by more than one result. For example, consider the query "peru facts", the user expects to see many good results to collecting information about peru. For informational queries, the novelty and redundancy concerns are important, and user does not know in advance which document has exact information. User reviews the result set of informational query in depth.

In order to process the informational queries, the user's intent and document classification techniques are used.



```
DiversityIQ Algorithm
("Rank document to maximize Equation 3*)
R ← θ
D ← All relevant documents
While IRI < n
        d ← AVERMAX (ΔE (d|R, D))
        R ← R U {d}
        D ← D {d}
```

Fig. 20 Algorithm diversity—IQ [44]

User's intent has perfect knowledge of user's interest. User's intent and document classification are used to find relevant subtopics, and the possibility of user's concern in all subtopics. This helps in producing an ordered set of documents, where an average user finds sufficient relevant documents.

In terms of diversification algorithms, diversity-IQ and IA-Select algorithms are used for this type of queries. The difference between the two algorithms is that diversity-IQ uses both user's intent and document classification; however, IA-Select is only based on document classification. The experimental results presented in the literature show that diversity-IQ performs better as compared to IA-Select in terms of finding the subtopics and classification of documents.

In future, the processing of such queries can be improved by conducting better classification of documents. Particularly, there is a need to work on the documents which belong to more than one class.

8 Discussion and future directions

This study presents a survey of the search result diversification techniques. It has been figured out that the problem of search result diversification has been addressed based on different types of user queries, each type of query is processed in a different manner so as to get the relevant and diversified results. Furthermore, the survey reveals that there are few diversification algorithms which are customized based on the type of the query. This work presents a classification of existing search result diversification based on the types of the queries, and it also provides a link of existing algorithms on different diversification methods identified in the proposed taxonomy.

Based on the analysis of the surveyed literature, it was observed that the problem of search result diversification has been addressed in the following different classes of queries: ambiguous query; unambiguous but underspecified query; geo-referenced query; multi-domain query; and informational query.

In short, the literature reveals that there exist a few diversification algorithms. Some people have used the baseline algorithms as they are, whereas others have provided variants of a baseline algorithm so as to customize them based on the requirements of the problem. Table 1 presents the widely used diversification algorithms and their brief description.

A principal benefit of this study is that, to the best of our knowledge, this is the first effort to compile all the work pertaining to the search result diversification. Furthermore, the output of this work can be useful for the information retrieval systems in general, search engine development and improvement in particular. They can map the input queries to the appropriate query class as defined in the proposed taxonomy and thus can figure out the most appropriate diversification technique to resolve a query. Lastly, it also provides future research directions and discusses evaluation measures used in search result diversification.

Table 1 Diversification algorithms and their brief description

Algorithm	Description
XQuAD	XQuAD algorithm is specially designed for subquery generation
IA-Select	IA-Select is specially designed for the methods which utilize subtopics, where the documents and queries are classified according to the identified subtopics
Opt-Select	Opt-Select is specially designed for manipulating the information extracted from query log
MMR	The MMR diversification algorithm uses the concept of text-based similarity measure such as the vector space model

8.1 Diversification algorithms

A diversification algorithm takes the top relevant results as input and processes them to produce relevant as well as diversified result sets. The diversification algorithms use a diversity measure to compute the difference between the items in the result set, whereas the dataset already possesses the relevance measure which reflects the similarity between the query and the documents. A diversification objective function is an integral part of a diversification algorithm. It incorporates both the relevance measure and the diversity measure to compute a diversified result set.

Different types of diversification algorithms are used for computing diversified results. IA-Select, xQuAD, MMR, and Opt-Select are some widely used algorithms.

XQuAD algorithm is specially designed for subquery generation, and it performs very well for the methods which incorporate subquery generation. IA-Select is specially designed for the methods which utilize subtopics, where the documents and queries are classified according to the identified subtopics. The experiments show that this method does not perform very well for the documents which belong to more than one category.

Opt-Select is especially designed for manipulating the information extracted from query log. Submission of ambiguous queries in the past is discovered through query log, which in turn helps to cover different understandings of the query.

The MMR diversification algorithm uses the concept of text-based similarity measure such as the vector space model. In MMR, the suggested set of candidate queries is retrieved from the query log.

Table 2 Types of queries and corresponding diversification algorithms

Type of query	Diversification algorithms
Ambiguous queries	Opt-Select
	Portfolio model
	IA-Select
	xQuAD
Unambiguous but under specified queries	PxQuAD
	PIA-Select
	MMR
Multi-domain queries	MMR
Geo-referenced queries	PBMMR
	SPP
Informational queries	Diversity-IQ

Table 2 relates the types of queries with the diversification algorithms studied in this research work. We can see that ambiguous queries are addressed based on many different diversification algorithms.

8.2 Diversity-aware evaluation measures and datasets

Relevance and novelty are two basic measures to evaluate the diversity among the results. *Relevance* involves relatedness of a result to the given query, whereas *novelty* reflects the measure of the involvement of different object categories in the result set. As an example, consider a query "windows" which involves *relevant* documents with many different perspectives. Covering all perspectives, e.g., room window, Microsoft windows with different documents represent the *novelty*, which in turn leads to diversification of search results.

In order to evaluate the effectiveness of discussed diversity-aware search approaches, the researchers have introduced new measures in the domain of *Information Retrieval*. Normalized discounted cumulative gain (NDCG) is a classical measure used for evaluating an information retrieval system's efficiency in terms of non-binary relevance, which has been customized for evaluating diversification as alpha-NDCG [48] and NDCG-IA. The NDCG-IA depends upon the intent distribution and on the intent-specific NDCG, whereas alpha-NDCG is used to evaluate the suptopics based on their relevance to the query and ensures the diversification of results based on already reported results. Along with NDCG-IA and alpha-NDCG, MAP-IA and MRR-IA [49] are other common metrics for user's intent. They consider ambiguous queries, which belong to different categories. They take into account the "popularity" of each query's category, for example consider the query "Jaguar" the car sense might be more prominent than the animal. Eventually, they help in the identification of the most relevant user intent for such ambiguous queries.

In the structured environment, such as a relational database system in an enterprise, the evaluation of the results is generally conducted based on the comparison of the computed result with the "optimal" result. In short, most of these metrics intend to incorporate both relevance and novelty in the result set. There exists a trade-off between the two which helps finding the relevant results while incorporating the user intent.

8.2.1 *Datasets for diversity-aware search*

Different types of datasets have been used for diversity-aware search. Many researchers used Wikipedia disambiguation pages for the evaluation of their work [2]. Text Retrieval Conference (TREC) is used for topics and a list of subtopics. Structured database is used for database-like search task. Open Directory Project (ODP) is used as a taxonomy to classify results.

It was observed that earlier work in the domain of search result diversification was evaluated based on nonstandard datasets. Therefore, in order to achieve diversity in result set in TREC 2009, the new "Diversity Task" started [50]. It was also noticed that in most cases, two main types of dataset have been used: classical textual documents to be ranked by TREC-like task, and structured dataset is used for database-like search task. In both cases, the goal was to provide the user with a smaller set of relevant and diverse results.

8.3 Future directions

There are several possible future directions in the area of search result diversification. One possible dimension is that there exists no specific diversification algorithm for queries such as "downloading software", and "watching movies". This will not only increase the coverage of the search result diversification but will also help in evolving the proposed taxonomy. The proposed taxonomy can be utilized by the information retrieval systems to map the user query onto a specific type of query class. This will certainly help in identifying and applying the most appropriate diversification algorithm for search result diversification with respect to the input query, which in turn will help in producing better quality results.

The ambiguous queries are generally processed using the query log, and the methods used to resolve such queries collect the statistics about most frequently accessed documents related to a query. The results of such methods can be improved by extracting more useful statistics about the documents, or by incorporating probabilistic measures over the gathered statistics.

In future, there is a need to classify the queries in such a way that the queries which can benefit from a search result diversification approach are distinguished from the ones which do not require search result diversification. To this end, the queries which involve the subtopic generation can be considered. For such queries, it is pertinent to identify the relative importance of each subtopic and then involve the most important subtopics in the process of producing a diversified result set.

Multi-domain queries have multiple linked concepts, and generally, the data for each concept are obtained from a different data service, which in turn exposes the data like a relational data model. Similarly, in future there is a need to develop new diversification approaches for multi-domain query, which may involve the relative importance of a data source and may also involve semantics to get more reasonably diversified result set.

Diversity and personalization can be joined in different ways, beyond which there exists a wide room of future research. For instance, there is a need to work on the exaggerated use of user's search history for personalizing resultant diversification.

Informational queries get relevant results by using the classification of documents. This can benefit from better classification of documents, particularly for the documents which belong to more than one class. Lastly, there is also a need to work on tackling the possible presence of uncertainty in the data while applying diversification techniques, particularly for geo-referenced queries.

Acknowledgments We are grateful to the reviewers of this article for their constructive and valuable review comments, which helped us to improve this article in many different ways.

References

1. Giunchiglia F (2006) Managing diversity in knowledge. Advances in applied artificial intelligence. Springer, Berlin, Heidelberg
2. Gollapudi S, Sharma A (2009) An Axiomatic Approach for Result Diversification. In: The international world wide web conference committee (IW3C2), ACM, Madrid
3. Zhai CX, Cohen WW, Lafferty J (2003) Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In: SIGIR'03 proceedings of the 26th annual international ACM SIGIR conference on research and development in information retrieval, ACM, New York, NY
4. Clough P, Sanderson M, Abouammoh M, Navarro S, Paramita M (2009) Multiple approaches to analysing query diversity. In: SIGIR'09 proceedings of the 32nd international ACM SIGIR conference on research and development in information retrieval, ACM, New York, NY
5. Jain A, Sarda P, Haritsa JR (2004) Providing diversity in K-nearest neighbor query. Advances in Knowledge Discovery and Data Mining. Springer, Berlin
6. Minack E, Demartini G, Nejdl W (2009) Current approaches to search result diversification. In: Proceedings of the first international workshop on living web, collocated with the 8th international semantic web conference (ISWC 2009), s.n., Washington D.C.
7. Sparck-Jones K, Robertson SE, Sanderson M (2007) Ambiguous requests: implications for retrieval tests, systems and theories, University of Cambridge ACM SIGIR, , s.n., UK, pp 8–17
8. Capannini G, Nardini FM, Perego R, Silvestri F (2011) Efficient diversification of web search results, VLDB Endowmen, Seattle, Washington
9. Rafiei D, Gollapudi S, Halverson A, Ieong S (2010) Diversifying web search results. In: the international world wide web conference committee (IW3C2), ACM, North Carolina
10. Zheng W, Wang X, Fang H, Cheng H (2011) An exploration of pattern-based subtopic modeling for search result diversification, JCDL, New York, NY. ISBN: 978-1-4503-0744-4
11. Santos RLT, Peng J, Macdonald C, Ounis I (2010) Explicit search result diversification through sub-queries. In: European conference, Springer, Berlin, pp 87–99
12. Capannini G, Nardini FM, Perego R, Silvestri F (2011) Efficient diversification of search results using query logs. In: WWW'11 Proceedings of the 20th international conference companion on World wide web, New York, NY, ACM
13. Boldi P, Bonchi F, Castillo C, Vigna S (2009) From "Dango" to "Japanese Cakes":Query reformulation models and patterns. WI'09. IEEE CS Press
14. Song R, Luo Z, Wen J-R, Y Yu (2007) Identifying ambiguous queries in web search. In: WWW'07 proceedings of the 16th international conference on world wide web, ACM, New York, NY
15. Richardson M, Dominowska E, Ragno R (2007) Predicting clicks: estimating the click-through rate for new ads.. In Proceedings of the 16th international world wide web conference(WWW-2007)
16. Vargas S, Santos RLT, Macdonald C, Ounis I (2013) Selecting effective expansion terms for diversity. In: OAIR'13 proceedings of the 10th conference on open research areas in information retrieval, France : Le Centre De Hautes Etudes Internationales D'Informatique Documentaire, Paris
17. Baeza-Yates R, Hurtado C, Mendoza M (2004) Query recommendation using query logs in search engines. In: EDBT'04 proceedings of the 2004 international conference on current trends in database technology, Springer, Berlin, Heidelberg
18. Santos RLT, Macdonald C, Ounis I (2010) Exploiting query reformulations for web search result diversification. In: WWW'10 proceedings of the 19th international conference on world wide web, ACM, New York, NY
19. Vallet D, Castells P (2012) Personalized diversification of search results. *In:* SIGIR'12 proceedings of the 35th international acm sigir conference on research and development in information retrieval, ACM, New York, NY
20. Vallet D, Castells P (2011) On diversifying and personalized web search, SIGIR'11, ACM, Beijing
21. Vallet D, Cantador I, Joemon M. Jose (2010) Personalizing web search with folksonomy-based user and document profiles. ECIR'2010 Proceedings of the 32nd European conference on Advances in Information Retrieval, Springer, Berlin
22. Sydow M, Ciesielski K, Wajda J (2011) Introducing diversity to log-based query suggestions to deal with underspecified user queries. In proceeding of: security and intelligent information systems—international joint conferences, DBLP, Warsaw
23. Radlinski F, Dumais S (2006) Improving personalized web search using result diversification. In: SIGIR'06 proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval, ACM, New York, NY

24. Speretta M, Gauch S (2005) Personalized search based on user search histories. In: Web intelligence, 2005, proceedings the 2005 IEEE/WIC/ACM International Conference, SPEC, USA
25. Vallet, D (2011) Crowdsourced evaluation of personalization and diversification techniques in web search. In: Proceedings of the SIGIR 2011 workshop on text retrieval conference (TREC)
26. Micarelli A, Gasparetti F, Sciarrone F, Gauch S (2007) Personalized search on the world wide web. In: The adaptive web Springer, Berlin
27. Bozzon A, Brambilla M, Ceri S, Fraternali P Liquid query: multidomain exploratory search. In: International world wide web conference, ACM, Raleigh, North Carolina, 2010
28. Ceri S, Brambilla M (2011) Search computing trends and developments. Springer Science & Business Media, Berlin, Heidelberg
29. Brambilla M, Brambilla M, Fraternali P, Tagliasacchi M (2011) Diversification for multi-domain result sets. In: CIKM'11 proceedings of the 20th ACM international conference on Information and knowledge management, ACM, New York, NY
30. Vee E, Srivastava U, Shanmugasundaram J, Bhat P (2008) Efficient computation of diverse query results. In: ICDE'08 proceedings of the 2008 IEEE 24th international conference on data engineering, IEEE Computer Society, Washington, DC
31. Martinenghi D, Tagliasacchi M (2010) Proximity rank join. In: Proceedings of the VLDB endowment
32. Demidova E, Fankhauser P, Zhou X, Nejdl W (2010) DivQ: diversification for keyword search over structured databases. In: SIGIR'10 proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, ACM, New York, NY
33. Carbonell J, Goldstein J (1998) The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: SIGIR'98 proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. vol ACM, s.n., New York, NY
34. Chen Y-Y, Suel T, Markowetz A Efficient query processing in geographic web search engines. In: SIGMOD'06 proceedings of the 2006 ACM SIGMOD international conference on management of data, ACM, New York, NY, 2006
35. Cong G, Jensen CS, Wu D (2009) Efficient retrieval of the top-k most relevant spatial web objects. Proc VLDB endow 2(1):337–348
36. Fraternali P, Martinenghi D, Tagliasacchi M (2012) Top-k bounded diversification. In: SIGMOD'12 proceedings of the 2012 ACM SIGMOD international conference on management of data, ACM, New York, NY
37. Schnaitter K, Polyzotis N (2008) Evaluating rank joins with optimal cost. In: PODS'08 proceedings of the twenty-seventh ACM SIGMOD-SIGACT-SIGART symposium on principles of database systems, ACM, New York, NY
38. Abid A, Tagliasacchi M (2013) Provisional reporting for rank joins. J Intell Inf Syst 40(3):479–500
39. Berchtold S, Ertl B, Keim DA, Kriegel H-P, Seidl T (1998) Fast nearest neighbor search in high-dimensional space. In: Proceedings of the 14th international conference on data engineering, Orlando, 23–27 Feb 1998
40. de Berg M, Cheong O, van Kreveld M, Overmars M (2008) Computational geometry: algorithms and applications. Springer, Netherlands
41. Soliman MA, Ilyas (2011) Ranking with uncertain scoring functions: semantics and sensitivity measures. In: SIGMOD'11 proceedings of the 2011 ACM SIGMOD international conference on management of data, ACM, ON, Canada. New York, NY
42. Bhatia S, Mitra P, Brunk C (2012) A query classification scheme for diversification. DDR'12, s.n., Seattle, WA, 2012
43. Lee U, Liu Z, Cho J (2005) Automatic identification of user goals in web search. In: WWW'05 proceedings of the 14th international conference on world wide web, ACM, New York, NY
44. Welch MJ, Cho J (2011) Search result diversity for informational queries. In: The international world wide web conference committee (IW3C2), ACM, Hyderabad
45. Qiu F, Liu Z, Cho J (2005) Analysis of user web traffic with a focus on search activities. In: International workshop on the web and databases
46. Blei DM (2012) Probabilistic topic models. Commun ACM 55(4):77–84
47. Blei DM, Ng AY (2003) Michael I Latent dirichlet allocation. J mach learn res 3:993–1022 **(s.n., Jordan)**
48. Clarke CLA, Kolla M, Cormack GV, Vechtomova O (2008) Novelty and diversity in information retrieval evaluation. In: SIGIR'08 proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval, ACM, New York, NY
49. Agrawal R, Gollapudi S, Halverson A, Ieong S (2009) Diversifying search results. In: WSDM'09 proceedings of the second ACM international conference on web search and data mining, ACM, New York, NY
50. Zhai C, Cohen WW, Lafferty J (2003) Beyond independent relevance: methods and evaluation. SIGIR-2003, ACM, Toronto