Inf Retrieval (2012) 15:433–457
DOI 10.1007/s10791-011-9178-4

Coverage-based search result diversification

Wei Zheng · Xuanhui Wang · Hui Fang · Hong Cheng

Received: 18 April 2011 / Accepted: 28 October 2011 / Published online: 17 November 2011

Abstract Traditional retrieval models may provide users with less satisfactory search experience because documents are scored independently and the top ranked documents often contain excessively redundant information. Intuitively, it is more desirable to diversify search results so that the top-ranked documents can cover different query subtopics, i.e., different pieces of relevant information. In this paper, we study the problem of search result diversification in an optimization framework whose objective is to maximize a coverage-based diversity function. We first define the diversity score of a set of search results through measuring the coverage of query subtopics in the result set, and then discuss how to use them to derive diversification methods. The key challenge here is how to define an appropriate coverage function given a query and a set of search results. To address this challenge, we propose and systematically study three different strategies to define coverage functions. They are based on summations, loss functions and evaluation measures respectively. Each of these coverage functions leads to a result diversification method. We show that the proposed coverage based diversification methods not only cover several state-of-the-art methods but also allows us to derive new ones. We compare these methods both analytically and empirically. Experiment results on two standard TREC collections show that all the methods are effective for diversification and the new methods can outperform existing ones.

W. Zheng (✉) · H. Fang
University of Delaware, 209 Evans Hall, Newark, DE 19716, USA
e-mail: zwei@udel.edu

H. Fang
e-mail: hfang@udel.edu

X. Wang
Yahoo! Labs, 4401 Great America Parkway, Santa Clara, CA 95054, USA
e-mail: xhwang@yahoo-inc.com

H. Cheng
The Chinese University of Hong Kong, William M. W. Mong Engineering Building,
Shatin, NT, Hong Kong
e-mail: hcheng@se.cuhk.edu.hk

Keywords Information retrieval · Diversification · Coverage · Subtopic

1 Introduction

Traditional retrieval models rank documents based on only their relevance scores and ignore the redundancy among the returned documents. As a result, the top ranked documents may contain the same piece of relevant information. It has been noticed that a large fraction of search queries are short and thus ambiguous or under-specified (Clarke et al. 2009a; Radlinski et al. 2009). For these queries, the targeted information for the same query can be quite different given different users. Search results covering different pieces of relevant information (i.e., subtopics of a query) are less risky and more desirable because they can provide diversified information that satisfies different information needs of users. For example, different users issuing the same query "java" may look for different information, such as java programming language and java coffee. A user searching for "cloud computing" may want to conduct a survey and learn different research topics related to cloud computing. It is clear that search result diversification can benefit ambiguous queries, under-specified queries and exploratory queries in general which account for a large portion of search queries (Clarke et al. 2009a; Radlinski et al. 2009; White and Roth 2009).

Search result diversification has recently attracted a lot of attentions (Clarke et al. 2009a; Radlinski et al. 2009; Zhai et al. 2003; Macdonald et al. 2011). The goal of result diversification is to return a list of documents which are not only relevant to a query but also cover many subtopics of the query. Here a query subtopic corresponds to a representative information need associated with the query, which is also referred to as a nugget (Clarke et al. 2008) or query aspect (Clarke et al. 2009b) in previous work. Using query subtopics for result diversification has received much attention. In particular, most of the commonly used evaluation measures, including α-*nDCG* (Clarke et al. 2008), *Precision-IA* (Agrawal et al. 2009), *NRBP* (Clarke et al. 2009) and *ERR-IA* (Chapelle et al. 2009), are all based on the coverage of query subtopics in the documents. Moreover, a few recent studies tried to diversify search results explicitly based on the query subtopics (Agrawal et al. 2009; Carterette and Chandar 2009; Santos et al. 2010b, c; Yin et al. 2009). However, none of existing studies has *systematically* studied and compared different subtopic-based diversification methods. Thus, the underlying commonalities of these methods are not well understood, and there is no guidance that can be used to derived new and possibly more effective diversification methods.

In this paper, we study the search result diversification problem in a coverage-based optimization framework. Following previous studies, we define the optimization objective function as a linear combination of relevance and diversity scores and then use a greedy algorithm to iteratively select results to maximize the object function. In particular, we propose to model the diversity scores of search results based on their coverage of query subtopics, i.e., how much relevant information of query subtopics is contained in the documents. The diversity score is directly related to the relevance of these results with respect to the subtopics and the importance of these subtopics. Moreover, with the assumption that all the query subtopics are independent, the diversity score is computed as the weighted sum of its coverage for every query subtopic. Thus, the key challenge is how to define appropriate coverage functions given a query subtopic and a set of search results. To address this challenge, we explore three different strategies. The first summation-based strategy is to compute the coverage score of a document set by summing up those of individuals in the set. The second strategy defines a coverage function through a loss function (e.g., squared loss) over a set of selected documents. The last one is to derive

coverage functions based on a few commonly used evaluation measures. With these three strategies, different variants of coverage functions are defined and each of them can lead to a potentially different coverage-based diversification method. We analyze the similarity among these methods and select five representative ones to study in this paper. Among the derived diversification methods, two of them are variants of some existing diversification methods (Agrawal et al. 2009; Santos et al. 2010c; Yin et al. 2009) and three of them are new diversification methods that have not been studied before. We first analytically compare the characteristics of these five methods from the following aspects: *diminishing return, favoring diversity*, and *novelty emphasis*, and then conduct experiments over two standard TREC collections. Experimental results show that most of the diversification methods derived in the optimization framework are effective. Furthermore, one of the new derived methods, i.e., *SQR*, can consistently outperform the state-of-the-art methods over different collections and with different retrieval models.

The rest of the paper is organized as follows. We discuss the related work in Sect. 2. We review the main idea of optimization framework for result diversification in Sect. 3. We then propose and study different coverage functions and corresponding diversification methods in Sect. 4. We analytically compare these diversification methods in Sect. 5 and empirically compare their results in Sect. 6. Finally, we conclude in Sect. 7.

2 Related work

The study of search result diversification can be traced back to early sixties (Goffman 1964). Since then, many studies have tried to rank documents based on not only relevance but also diversity (Agrawal et al. 2009; Boyce 1982; Carbonell and Goldstein 1998; Carterette and Chandar 2009; Chen and Karger 2006; Gollapudi and Sharma 2009; McCreadie et al. 2009; Radlinski et al. 2009; Radlinsk and Dumais 2006; Santos et al. 2010c; Yin et al. 2009; Yue and Joachims 2008; Zhai et al. 2003). Roughly speaking, the proposed methods can be classified into two categories (Santos et al. 2010b, c).

The first category implicitly models diversity through the relations among documents in order to minimize the redundant information among the selected documents (Carbonell and Goldstein 1998; Chen and Karger 2006; Craswell et al. 2009; Demidova et al. 2010; Gollapudi and Sharma 2009; Santos et al. 2010b;Yue and Joachims 2008; Zhai et al. 2003). Carbonell and Goldstein (1998) proposed the maximal marginal relevance (*MMR*) ranking strategy to balance the relevance and the redundancy. Motivated by this work, Zhai et al. (2003) combined both relevance and novelty in the statistical language modeling framework. Chen and Karger (2006) presented a sequential document selection algorithm to optimize an objective function that aims to find at least one relevant document for all users. Yue and Joachims (2008) treated the diversification as a process to learn the function of choosing the optimum set of diversified documents, which is a learning problem. Their learned function can sequentially select documents which cover maximally distinct words. Craswell et al. (2009) removed the redundant documents based on the host information of the documents. Gollapudi and Sharma (2009) proposed an axiomatic approach to characterize the problem of result diversification and studied several redundancy functions in their axiomatic framework. The main difference among different methods is how to model the redundancy between a new document and the previous selected documents without an explicit modeling of query subtopics.

The second category of search result diversification explicitly model diversity among documents through their relation to the subtopics of the queries (Agrawal et al. 2009; Carterette and Chandar 2009; Radlinsk and Dumais 2006; Santos et al. 2010a, b, c; Yin

et al. 2009). Subtopics are often identified using topic modeling (Carterette and Chandar 2009), existing taxonomy (Agrawal et al. 2009), query suggestions (Santos et al. 2010b) or frequent patterns (Zheng and Fang 2010). Most of the state-of-the-art diversification methods often formalize the problem in an optimization framework, where the objective function is defined based on the combination of relevance and diversity scores. Since it is an NP-hard problem to select an optimum document set in general, a greedy algorithm is often used to iteratively select documents. In particular, Carterette and Chandar (2009) proposed a probabilistic set-based approach that maximizes the likelihood of capturing all of the query subtopics. Agrawal et al. (2009) formalized the problem as the one that maximizes average user satisfaction based on a classification taxonomy over queries and documents. Santos et al. (2010b, c) proposed a probabilistic framework that estimates the diversity based on the relevance of documents to query subtopics and the importance of query subtopics. Yin et al. (2009) derived a diversification method using the language modeling approach. Since these existing diversification methods were proposed and studied in different work, the connection of these methods remains unclear. It is difficult to analytically compare these methods although there are empirical comparisons among them (Zheng and Fang 2011). Moreover, there is no systematic way of developing new diversification methods. Finally, almost all of the existing diversification methods are based on probabilistic models, and it remains unclear how to diversify search results for non-probabilistic retrieval models.

Developing effective evaluation measures has also attracted a lot of attention. Most commonly used evaluation measures, including *Precision-IA* (Agrawal et al. 2009), *ERR-IA* (Chapelle et al. 2009), *α-nDCG* (Clarke et al. 2008) and *NRBP* (Clarke et al. 2009b), are all based on the coverage of query subtopics. We review them briefly in this section.

Intent-aware precision at retrieval depth k (*Precision-IA@k*) is based on a weighted average of precision at depth k across different subtopics (Agrawal et al. 2009; Clarke et al. 2009a). The *Precision-IA@k* in a query q can be computed as

$$Precision-\mathrm{IA@k} = \sum_{s \in S(q)} \mathrm{weight(s,q)} \cdot \frac{1}{k} \sum_{j=1}^{k} \mathrm{r(d_j, s)}, \tag{1}$$

where $S(q)$ is the subtopic set of query q, *weight*(s, q) is the importance weight of a subtopic s in the query and is computed as $\frac{1}{|S(q)|}$, k is the depth and $r(s, d_j)$ is the relevance judgment of the document d_j to the subtopic s. *Precision-IA@k* uses the binary relevance score as $r(s, d_j)$.

Intent-aware expected reciprocal rank at retrieval depth k (*ERR-IA@k*) uses cascade-style user browsing model to estimate the expected reciprocal length of time for the user to find a relevant document. It is computed as follows (Chapelle et al. 2009):

$$ERR-\mathrm{IA@k} = \sum_{s \in S(q)} \mathrm{weight(s,q)} \sum_{j=1}^{k} \frac{\mathrm{r'(d_j, s)}}{j} \prod_{i=1}^{j-1} (1 - \mathrm{r'(d_i, s)}), \tag{2}$$

where $r'(d_j, s)$ is the probability of relevance mapped from relevance grade of d_j to s.

Another commonly used evaluation measure is *α-nDCG* (Clarke et al. 2009a). It integrates the novelty of subtopic into normalized discounted cumulative gain measure. It is computed as:

$$\alpha-\mathrm{nDCG@k} = \frac{\alpha-\mathrm{DCG@k}}{\alpha-\mathrm{DCG'@k}}, \tag{3}$$

where

$$\alpha-\mathrm{DCG@k} = \sum_{\mathrm{j}=1}^{\mathrm{k}} \frac{\sum_{\mathrm{s}\in\mathrm{S(q)}} \mathrm{r(d_j, s)}(1-\alpha)^{\sum_{\mathrm{i}=1}^{\mathrm{j}-1} \mathrm{r(d_i,s)}}}{\log_2(1+\mathrm{j})} \tag{4}$$

and α-*DCG*′*@k* is the maximum value of α-*DCG@k* given the ideal order of the returned document list.

Novelty- and Rank-Biased Precision (*NRBP*) is a measure that combines user browsing model, intent aware and α-*nDCG*. The function is as follows:

$$NRBP = \frac{1-(1-\alpha)\beta}{N} \sum_{j=1}^{\infty} \beta^{j-1} \sum_{s\in S(q)} r(d_j, s)(1-\alpha)^{\sum_{i=1}^{j-1} r(d_i,s)} \tag{5}$$

where α is the parameter to describing user's ability to judge the relevance of document given the subtopic and β is the parameter to estimate user's interest in reading more documents after finding one relevant document.

Intuitively, these evaluation measures can provide guidance on how to derive subtopic-based diversification methods. As mentioned in a recent study (Chapelle et al. 2009), a state-of-the-art diversification method, i.e., *IA-SELECT* (Agrawal et al. 2009), is similar to a standard evaluation measure, i.e., *ERR-IA* (Chapelle et al. 2009). Sakai and Song (2011) compared the existing evaluation measures and analyzed the advantages of each method in queries with different types of subtopics. They found that α-*nDCG* is one of the best evaluation measures in the existing measures. However, to our best knowledge, none of the existing work studied how to derive new diversification methods based on the evaluation measures for diversity.

Our work is similar to previous studies (Agrawal et al. 2009; Santos et al. 2010b) in the sense that we also formulate the problem as an optimization problem and use a greedy algorithm to select documents. However, our framework is more general that provides a systematically way of deriving and analyzing new diversification methods. Under the guidance of the framework, we define different functions measuring the coverage of the document on subtopics and systematically derive five diversification methods, three of which are new diversification methods. We then analytically and empirically compare these methods.

3 An optimization framework for result diversification

The goal of result diversification is to return a set of relevant search results that can cover diverse pieces of relevant information. The problem is often formulated as an optimization problem that aims to maximize an objective function related to both the *relevance* and *diversity* of the search results (Agrawal et al. 2009; Carbonell and Goldstein 1998; Santos et al. 2010b, c; Zhai et al. 2003).

Formally, given a query q, a set of documents $\mathcal{D}$ and an integer k, the goal is to find D, i.e., a subset with k documents from the document set, that can maximize the following objective function:

$$G(D) = \lambda \cdot rel(q, D) + (1-\lambda) \cdot div(q, D), \tag{6}$$

where $D \subseteq \mathcal{D}$. The objective function is based on both $rel(q, D)$, which measures the relevance score of document set D with respect to the query q, and $div(q, D)$, which measures the diversity score of D for q. It is similar to the ideas of existing methods

(Carbonell and Goldstein 1998; Santos et al. 2010b, c). $\lambda \in [0, 1]$ is a parameter that controls the tradeoff between diversity and relevance. When $\lambda = 1$, the optimization goal is to select top k documents ranked based on their relevance scores, which is consistent with the traditional retrieval models.

Finding the optimal solution for the above problem is in general NP-hard since it can be reduced from Maximum Coverage Problem (Agrawal et al. 2009). Fortunately, if $G(D)$ is a submodular function, a greedy algorithm which sequentially selects a document that maximizes the marginal gain of the submodular function can achieve $(1 - \frac{1}{e})$ approximation of the optimal solution and has been shown to be almost optimal in practices (Agrawal et al. 2009; Khuller et al. 1999; Leskovec et al. 2007).

Specifically, the greedy algorithm starts with an empty document set $D = \varnothing$, and then iteratively selects a local optimal document, which is defined as follows (Agrawal et al. 2009; Santos et al. 2010b):

$$d^* = \arg \max_{d \in \mathcal{D} \setminus D} (G(D \cup \{d\}) - G(D)) \tag{7}$$

The local optimal document is then added to document set D until the number of documents in D is k. The order that the documents are selected gives us a ranked list and it only depends on the definition of the objective function.

We now discuss how to define an objective function. The objective function is related to $rel(q, D)$ and $div(q, D)$, as shown in (6). $rel(q, D)$ measures how much relevant information is contained in the document set with respect to the query. One possible way of computing $rel(q, D)$ is

$$rel(q, D) = \sum_{d \in D} rel(q, d), \tag{8}$$

where $rel(q, d)$ is the relevance score of document d for query q and can be computed using any existing retrieval functions.

Existing studies mainly differ in how to compute $div(q, D)$, which measures the diversity score of document set D with respect to query q. One strategy is to compute the diversity score based on the redundancy of the document set D (Carbonell and Goldstein 1998; Zhai et al. 2003). The diversity score is smaller when there is more redundant information in the document set. One major limitation of this approach is that the diversity is query-independent. Thus, the diversity score can be arbitrarily boosted by the non-relevant information of the search results and may not be used to effectively diversify relevant search results (Santos et al. 2010b). An alternative strategy is to compute the diversity score based on the query subtopics. (Note that query subtopics are also referred to as nuggets or query aspects in previous studies (Clarke et al. 2008; Santos et al. 2010b, c).) Most existing studies adopt probabilistic methods. For example, Agrawal et. al. (2009) proposed an objective function based on only the diversity score, which is estimated with the probability that the document set would satisfy the user who issues the query. The probabilities are estimated based on a classification taxonomy. Santos et. al. (2010b) proposed an objective function based on the relevance of documents to query subtopics and the importance of query subtopics in a probabilistic framework. However, it is difficult to derive more diversity functions using these methods and analytically compare different methods. It remains unclear whether it is feasible to define a diversity function using a more general approach that can be used to guide the derivation of new diversification functions.

4 Coverage-based diversification methods

In this paper, we aim to explore a general way of computing the diversity score, i.e., $div(q, D)$. In this section, we first describe the general idea of modeling diversity based on the *coverage* of query subtopics, and then propose three strategies for defining coverage functions. The coverage measures the relevant information of the query subtopics contained in the documents. After that, we explain how to derive the diversification methods based on the coverage functions.

4.1 General idea

Let $S(q)$ denote a set of subtopics of the query q. We propose to compute the diversity score as follows:

$$div(q, D) = \sum_{s \in S(q)} weight(s, q) \cdot cov(s, D), \tag{9}$$

where $weight(s, q)$ measures the importance of the subtopic s of the query q and $cov(s, D)$ measures the coverage of a specific subtopic s in the document set D. It assumes that the subtopics are covering different relevant information of the query and are independent of each other. Intuitively, the more subtopics that D covers and the more important that the covered subtopics are, the higher diversity score that D has. Our definition is consistent with existing methods (Santos et al. 2010b, c; Yin et al. 2009) and more general.

Given the coverage-based diversity function, we can re-write the objective function shown in (6) as follows:

$$G(D) = \lambda \cdot rel(q, D) + (1 - \lambda) \cdot \sum_{s \in S(q)} weight(s, q) \cdot cov(s, D). \tag{10}$$

As described in Sect. 3, it is NP-hard problem to find the optimum set of diversified documents. Given an objective function $G(D)$, we need to prove that the objective function is submodular in order to use the greedy algorithm to approximate the solution of the optimization problem. With a submodular function, the benefit of adding an element to a document set is not larger than adding the element to a subset of the document set. Thus, in order to prove that a function $G(D)$ is a submodular function with respect to D, we need to show that, for all sets $A, B \subseteq D$ such that $A \subseteq B$, and $d \in D \setminus B$, we have

$$(G(A \cup \{d\}) - G(A)) - (G(B \cup \{d\}) - G(B)) \geq 0.$$

As shown in (10), the objective function $G(D)$ is a linear combination of two components, i.e., $rel(q, D)$ and $cov(q, D)$. It is clear that the relevance score, i.e., $rel(q, D)$ in (8), is a submodular function with respect to D:

$$\begin{aligned} &(rel(q, A \cup \{d\}) - rel(q, A)) - (rel(q, B \cup \{d\}) - rel(q, B)) \\ &= \left(\sum_{d \in A \cup \{d\}} rel(q, d) - \sum_{d \in A} rel(q, d) \right) - \left(\sum_{d \in B \cup \{d\}} rel(q, d) - \sum_{d \in B} rel(q, d) \right) \\ &= 0. \end{aligned} \tag{11}$$

Since the linear combination of submodular functions is still a submodular function (Nemhauser et al. 1978), $G(D)$ is a submodular function if $cov(q, D)$ is a submodular function. In order to prove $cov(q, D)$ is a submodular, we need to show

$$(cov(s, A \cup \{d\}) - cov(s, A)) - (cov(s, B \cup \{d\}) - cov(s, B)) \geq 0.$$

In summary, in order to prove $G(D)$ is a submodular function with respect to D, for all sets $A, B \subseteq D$ such that $A \subseteq B$, and $d \in D \setminus B$, we need to show

$$DIFF(s, d, A) - DIFF(s, d, B) \geq 0, \tag{12}$$

where

$$DIFF(s, d, D) = cov(s, D \cup \{d\}) - cov(s, D). \tag{13}$$

Therefore, in each diversification method, we need to prove that (12) holds in order to use the greedy algorithm described in Sect. 3 that diversifies search results by iteratively selecting local optimal documents. According to (7), (10) and (13), the local optimal document d^* can be selected based on:

$$d^* = \arg \max_{d \in \mathcal{D} \setminus D} (\lambda \cdot rel(q, d) + (1 - \lambda) \cdot \sum_{s \in S(q)} weight(s, q) \cdot DIFF(s, d, D)). \tag{14}$$

4.2 Coverage functions

We now discuss how to define the coverage function, i.e., $cov(s, D)$, which measures how well a document set D covers the information of the query subtopic s. Intuitively, $cov(s, D)$ is related to the subtopic coverage of each document in the set, i.e., $cov(s, d)$, where $d \in D$. Furthermore, as discussed in the previous subsection, we require that $cov(s, D)$ should be a submodular function. Thus, the problem is how to combine the coverage of individual documents in D so that $cov(s, D)$ is submodular, i.e., the difference function defined in (13) satisfies the requirement shown in (12). We explore three set of methods to compute the coverage of D based on the coverage of documents in D, the coverage of documents that are not included in D, and following the idea of evaluation measures.

4.2.1 Summation-based coverage functions

A simple strategy of computing the coverage score of a document set is to sum up the coverage scores of its individual documents. We explore the following two ways of combining the individual coverage scores.

– **SUM:** It assumes that the coverage of a document set on a subtopic increases linearly with the coverage of each document on the subtopic. Therefore, we combine the coverage of each document in the document set by taking the summation over them.

$$cov_{SUM}(s, D) = \sum_{d \in D} cov(s, d) \tag{15}$$

Thus, we have

$$\begin{aligned} DIFF_{SUM}(s, d, D) &= cov_{SUM}(s, D \cup \{d\}) - cov_{SUM}(s, D) \\ &= cov(s, d). \end{aligned} \tag{16}$$

We prove that $cov_{SUM}(s, D)$ and its diversification function in (9) are submodular functions in Theorem 1 in the appendix.

– **LOG:** It is similar to *SUM* but with a log transformation to ensure the decrease of gain when adding a document covering the subtopic that has already been well covered. In (16), the increase of coverage on the subtopic s by adding the same document covering s is the same no matter whether the subtopic has been well covered by D, which may tend to rank redundant documents in the top of the results (Santos and Ounis 2011). However, the benefit of adding a document covering s should be smaller if the subtopic has already been well covered by D and this is desired from end users' viewpoints (Clarke et al. 2009a). We therefore propose *LOG* to solve the problem. The coverage of the document set grows sublinearly with the coverage of each document as follows:

$$cov_{LOG}(s, D) = \log\left(1 + \sum_{d \in D} cov(s, d)\right) \tag{17}$$

So, we have

$$\begin{aligned} DIFF_{LOG}(s, d, D) &= \log\left(1 + \sum_{d' \in D \cup \{d\}} cov(s, d')\right) - \log\left(1 + \sum_{d' \in D} cov(s, d')\right) \\ &= \log\left(1 + \frac{cov(s, d)}{1 + \sum_{d' \in D} cov(s, d')}\right), \end{aligned} \tag{18}$$

We prove that $cov_{LOG}(s, D)$ is a submodular function in Theorem 2.

4.2.2 Loss-based coverage functions

In the second strategy, we propose to define coverage functions of a document set based on the coverage of documents that are not included in the document set $cov(s, \overline{D})$. Without loss of generality, we assume the values of $cov(s, d)$ in each subtopic are normalized so that $cov(s, d)$ and $cov(s, D)$ are between 0 and 1.

– **PCOV:** We follow the idea of derivation based on probability model in *xQuAD* (Agrawal et al. 2009) to derive *PCOV* which stands for "probabilistic coverage".

$$cov_{PCOV}(s, D) = 1 - cov(s, \overline{D}) = 1 - \prod_{d \in D}(1 - cov(s, d)), \tag{19}$$

and

$$\begin{aligned} &DIFF_{PCOV}(s, d, D) \\ &= (1 - \prod_{d' \in D \cup \{d\}} (1 - cov(s, d'))) - (1 - \prod_{d' \in D} (1 - cov(s, d'))) \\ &= \prod_{d' \in D} (1 - cov(s, d')) - (1 - cov(s, d)) \prod_{d' \in D} (1 - cov(s, d')) \\ &= cov(s, d) \cdot \prod_{d' \in D} (1 - cov(s, d')), \end{aligned} \tag{20}$$

In fact, if $cov(s, d)$ is treated as the probability that document d is relevant to the query subtopic s, $cov_{PCOV}(s, D)$ can also be interpreted as the probability that at least one document from D is relevant to s (Agrawal et al. 2009; Santos et al. 2010b).

– **SQR:** *SQR* is a loss-based method which is not restricted to probability model and uses the squared loss to define coverage function.

$$cov_{SQR}(s, D) = 1 - \left(1 - \sum_{d \in D} cov(s, d)\right)^2 \tag{21}$$

Squared loss functions have been widely used in regression problems (Hastie et al. 2009). In our case, more generally, we can define the loss function based on any power γ.

$$cov_{POW}(s, D) = 1 - \left(1 - \sum_{d \in D} cov(s, d)\right)^{\gamma} \tag{22}$$

When $\gamma = 1$, the coverage function is the same with *SUM* where the coverage of a document set on a subtopic increases linearly with the coverage of each document. The power γ provides flexibility to model non-linear relationship of the coverage as described in Sect. 4.2.1. For any $\gamma \geq 1$, the defined function can be proved to be submodular. In our paper, we focus our study on *SQR* where $\gamma = 2$ and leave other settings as future work. For *SQR*, we have

$$\begin{aligned} &DIFF_{SQR}(s, d, D) \\ &= 1 - \left(1 - \sum_{d' \in D \cup \{d\}} cov(s, d')\right)^2 - \left(1 - (1 - \sum_{d' \in D} cov(s, d'))^2\right) \\ &= \left(1 - \sum_{d' \in D} cov(s, d')\right)^2 - \left(1 - cov(s, d) - \sum_{d' \in D} cov(s, d')\right)^2 \\ &= cov(s, d) \cdot \left(2 - 2 \cdot \sum_{d' \in D} cov(s, d') - cov(s, d)\right). \end{aligned} \tag{23}$$

We prove $cov_{SQR}(s, D)$ is a submodular function in Theorem 4.

4.2.3 Measure-based coverage functions

Another possible way of defining coverage functions is based on evaluation measures for diversity, since most of them are designed based on query subtopics as well. In particular, we study four commonly used measures, i.e., *Precision-IA*, *ERR-IA*, *α-nDCG* and *NRBP*. In the following, we assume that the relevance judgment $r(s, d)$ in (1)–(5) can be estimated using $cov(s, d)$, where $0 \leq cov(s,d) \leq 1$.

– **EVAL1:** Intent-aware precision at retrieval depth k (*Precision-IA@k*) is based on a weighted average of precision at depth k across different subtopics (Agrawal et al. 2009; Clarke et al. 2009a). Since the measure is a set-based measure, it is straightforward to follow the same intuition of the measure and define the coverage function as follows:

$$cov_{EVAL1}(s, D) = \frac{1}{|D|} \sum_{d \in D} cov(s, d).$$

Unfortunately, $cov_{EVAL1}(s,D)$ is not a submodular function as shown in Theorem 5. However, it is still interesting to see that dropping off $\frac{1}{|D|}$ could lead to a submodular function, which is the same as $cov_{SUM}(s, D)$.

Table 1 Coverage-based diversification methods

Names	Coverage functions	Diversification methods
SUM	$DIFF_{SUM}$ and $DIFF_{EVAL1}$	$\lambda \cdot rel(q,d) + (1-\lambda) \cdot \sum_{s \in S(q)} (weight(s,q) \cdot cov(s,d))$
LOG	$DIFF_{LOG}$	$\lambda \cdot rel(q,d) + (1-\lambda) \cdot \sum_{s \in S(q)} (weight(s,q) \cdot \log(1 + \frac{cov(s,d)}{1+\sum_{d' \in D} cov(s,d')}))$
PCOV	$DIFF_{PCOV}$ and $DIFF_{EVAL2}$	$\lambda \cdot rel(q,d) + (1-\lambda) \cdot \sum_{s \in S(q)} (weight(s,q) \cdot cov(s,d) \cdot \prod_{d' \in D} (1 - cov(s,d')))$
SQR	$DIFF_{SQR}$	$\lambda \cdot rel(q,d) + (1-\lambda) \cdot \sum_{s \in S(q)} (weight(s,q) \cdot cov(s,d) \cdot (2 - 2 \cdot \sum_{d' \in D} cov(s,d') - cov(s,d)))$
EVAL	$DIFF_{EVAL3}$	$\lambda \cdot rel(q,d) + (1-\lambda) \cdot \sum_{s \in S(q)} (weight(s,q) \cdot cov(s,d) \cdot (1-\alpha)^{\sum_{d' \in D} cov(s,d')})$

– **EVAL2:** Intent-aware expected reciprocal rank at retrieval depth k (*ERR-IA@k*) is a ranking list based measure, which can not be directly used to define the coverage function $cov(s, D)$ since D should be a set. However, this measure suggests a way for computing the contribution of the document ranked at $|D| + 1$, which is exactly what $DIFF(s, d, D)$ models. Thus, we can directly define $DIFF_{EVAL2}(s, d, D)$ based on the function of *ERR-IA@k*:

$$DIFF_{EVAL2}(s,d,D) = \frac{cov(s,d)}{|D|+1} \prod_{d' \in D} (1 - cov(s,d')). \tag{24}$$

The coverage function based on $DIFF_{EVAL2}$ is a submodular function as described in Theorem 6. We can drop off $\frac{1}{|D|+1}$ in $DIFF_{EVAL2}(s, d, D)$ since it does not affect the document selection in the greedy algorithm. It is the same as the $DIFF_{PCOV}$.

– **EVAL3:** α-*nDCG* and *NRBP* are also commonly used evaluation measures (Clarke et al. 2009a, 2008, 2009b). They are ranking-list based measures, and can be used to compute the coverage difference as follows:

$$DIFF_{EVAL3}(s,d,D) = cov(s,d) \cdot (1-\alpha)^{\sum_{d' \in D} cov(s,d')}, \tag{25}$$

where $k = |D|$, and α is a parameter. The larger α is, the more impact that $cov(s, d)$ has on the coverage score.

We prove that *EVAL3* is a submodular function in Theorem 7. As seen in (12) and (14), a diversification method is directly determined by $DIFF(s, d, D)$. Thus, even if we can not explicitly define a coverage function based on this measure, the derived $DIFF_{EVAL3}(s, d, D)$ is enough to be used to derive a new diversification method.

Plugging the *DIFF* functions [as shown in (16)–(25)] into (14), we can derive different diversification methods as shown in Table 1.

4.3 Discussions

The objective function in (10) has a few component, i.e., $rel(q, d)$, $weight(s, q)$ and $cov(s, d)$, that we need to discuss how to compute. These components can be instantiated

using different retrieval models. For example, $rel(q, d)$ and $cov(s, d)$ can be computed using any retrieval models by treating q or s as a query. In fact, this is one major advantage of the framework because the derived diversification methods are general and can be combined with any existing retrieval functions.

Since most existing diversification methods are based on probabilistic models, we now describe how to instantiate the component functions using language modeling approaches (Lafferty and Zhai 2001; Ponte and Croft 1998) and discuss the connections between our derived methods and the state-of-the-art methods.

In language modeling framework, the relevance score of a document d given query q, $P(d|q)$, is usually estimated as follows (Lafferty and Zhai 2001):

$$P(d|q) \propto P(q|d)P(d),$$

where $P(q|d)$ is the query likelihood (Zhai and Lafferty 2001) and $P(d)$ is the prior of d. Since $\sum_{s \in S(q)} weight(s, q)cov(s, d)$ is used to compute the diversity in our framework, one natural way of instantiating $weight(s, q)$ and $cov(s, d)$ in the probabilistic models is as follows:

$$weight(s, q) = P(s|q)$$
$$cov(s, d) = P(d|s)$$

where $P(s|q)$ is the probability that the subtopic s is relevant to q and $P(d|s)$ is the probability that d is relevant to s (Agrawal et al. 2009; Santos et al. 2010b). Both probabilities can be estimated in a similar way as $P(d|q)$. We compute these probabilities using Dirichlet method (Zhai and Lafferty 2001).

With these instantiations, it is interesting to see that the diversification method *SUM* is similar to the existing method *WUME* (Yin et al. 2009) and *PCOV* is similar to *IA-Select* and *xQuAD* methods (Agrawal et al. 2009; Santos et al. 2010b). The main differences between our methods and these existing methods are the subtopic extraction and component estimation. For example, we estimate $P(s|q)$ and $P(d|s)$ based on Dirichlet method (Zhai and Lafferty 2001). The existing methods estimate $P(s|q)$ based on the query suggestion scores from web search engines (Yin et al. 2009), relevance between s and q (Agrawal et al. 2009), popularity of subtopics in the collection (Santos et al. 2010b) and the coherence between retrieval results of query and subtopics (Santos et al. 2010b). They also used different similarity measures to estimate $P(d|s)$, i.e., BM25, DPH (Divergence From Randomness) model and language modeling (Agrawal et al. 2009, Santos et al. 2010b; Yin et al. 2009). The other difference between *PCOV* and *IA-Select* is that *IA-Select* does not consider the relevance score of the documents given the query in their diversification function while *PCOV* integrates the relevance score in (14).

5 Analytical comparison of diversification methods

In this section, we describe the desirable properties of the diversification method and analytically compare the proposed diversification methods based on these properties.

1. **Diminishing return**. Intuitively, if the document d covers the subtopics that have been better covered by previously selected documents in D, the gain of selecting this

document should be smaller. Thus, it is desirable to include the coverage of previously selected documents in the diversification methods. Among the five analyzed methods, *SUM* is the only function that ignores the relationship between a document d and the documents that have been selected, i.e., the documents in D. It cannot satisfy the *diminishing return* property. Thus, we predict that *SUM* diversification method performs worse than the other functions.

2. **Favoring diversity**. The underlying assumption of this property is that we should favor documents that cover more subtopics (Santos et al. 2010b). From our analysis, we find that *SQR* and *LOG* favor documents that cover more subtopics while the other functions may not have such a desirable property. This can be clearly seen in the following situation. When we start selecting the first document, i.e., $D = \emptyset$, *SQR* and *LOG* are strictly concave functions, so they favor documents that cover more subtopics. In the example of Table 2, q has 3 subtopics with equal weights. d_2 covers only one subtopic while d_1 covers two subtopics. Moreover, the degree of coverage is the same for these two documents, i.e., $cov(s_2, d_2) = cov(s_1, d_1) + cov(s_2, d_1)$. In the result of diversification methods, *Yes* means the methods always select the desired document on current position, *No* means it will not select the desired document, *Poss* means it will either select the desired document or other documents, and *Cond* means the method will select desired document if its parameter value satisfies the condition. With *SQR* and *LOG*, d_1 would be always selected first, which is clearly what we want because d_1 covers more subtopics. However, the other functions may select d_2 first.
3. **Novelty emphasis**. This property captures the intuition that we should favor documents with subtopics that are not well covered by the previously selected documents. Our analysis suggests that *SQR* and *EVAL* are more effective than the other functions in implementing this property. This can be seen in the same example in Table 2. Assume all the methods have selected d_1 and $D = \{d_1\}$. Intuitively, d_3 is more preferred than d_2 because d_3 covers a novel subtopic from the previously selected document d_1. The diversity functions would favor documents with higher $DIFF(q, d, D)$. For *SQR*,

Table 2 An illustrative example (top) for comparing coverage-based diversification methods and the ranking results (bottom) of diversification methods

Subtopics	$cov(s, d)$			$weight(s, q)$
	d_1	d_2	d_3	
s_1	0.1	0	0	0.33
s_2	0.1	0.2	0	0.33
s_3	0	0	0.18	0.33

Document ranking	Methods				
	SUM	*LOG*	*PCOV*	*SQR*	*EVAL*
Select d_1 first?	Poss.	Yes	Poss.	Yes	Poss.
Select d_3 (if already selected d_1)?	No	No	Poss.	Yes	Cond. ($\alpha > 0.651$)

$$
\begin{aligned}
diff_{SQR} &= DIFF(q, d_3, \{d_1\}) - DIFF(q, d_2, \{d_1\}) \\
&\propto cov(s_3, d_3) \cdot (2 - 2 \cdot cov(s_3, d_1) - cov(s_3, d_3)) \\
&- cov(s_2, d_2) \cdot (2 - 2 \cdot cov(s_2, d_1) - cov(s_2, d_2)) \\
&\propto (1 - \frac{cov(s_3, d_3) + cov(s_2, d_2)}{2}) \cdot (-\beta) + cov(s_2, d_2) \cdot cov(s_2, d_1)
\end{aligned}
$$

where $\beta = cov(s_2, d_2) - cov(s_3, d_3)$.

The requirement of selecting d_3 is $diff_{SQR} > 0$. According to the above equation, we can find that the requirement is:

- *SQR*: $\beta < \frac{cov(s_2,d_2) \cdot cov(s_2,d_1)}{1 - 0.5 \cdot (cov(s_3,d_3) + cov(s_2,d_2))}$

Similarly, we can get the requirement of selecting d_3 in other methods as follows:

- *SUM*: $\beta < 0$
- *LOG*: $\beta < cov(s_2, d_1) \cdot cov(s_3, d_3)$
- *PCOV*: $\beta < cov(s_2, d_1) \cdot cov(s_2, d_2)$
- *EVAL*: $\frac{cov(s_3,d_3)}{cov(s_2,d_2)} > (1 - \alpha)^{cov(s_2,d_1)}$

We assume that the values of $cov(s_2,d_2)$ and $cov(s_3,d_3)$ are unknown and compare the requirements in different methods. When $cov(s_3, d_3) > cov(s_2, d_2)$, all of these requirement are satisfied and all functions select d_3 first. When $cov(s_3, d_3) = cov(s_2, d_2)$, the methods *LOG*, *PCOV*, *EVAL* and *SQR* can select d_3 first while *SUM* may not. When $cov(s_3, d_3) < cov(s_2, d_2)$, the requirement of *SUM* is not satisfied. The upper bound of *EVAL* varies with different value of α. In the other methods, the upper bound of β in *SQR* is the largest, the upper bound in *PCOV* is smaller than that in *SQR* and the upper bound in *LOG* is smaller than that in *PCOV*. Given the data in Table 2, *SUM*, *LOG* will select d_2 before d_3, *PCOV* may select either d_2 or d_3, *EVAL* will select d_3 when $\alpha > 0.651$ and *SQR* will select d_3. Therefore, *SQR* and *EVAL* are more effective in favoring documents relevant to novel subtopics.

As we discussed in Sect. 4.3, *WUME* (Yin et al. 2009) is similar to *SUM*, and *IA-Select* (Agrawal et al. 2009) and *xQuAD* (Santos et al. 2010b) are similar to *PCOV*. These existing methods have the same properties with the corresponding methods proposed in this paper. The reason is that the properties listed above are only related to the relationships between different components in the functions while not related to the method of computing each component.

In summary, our analysis suggests that *SQR* is the most effective diversification method while *SUM* is the least effective one. Experiment results shown in Sect. 6 are consistent with our analysis.

6 Empirical comparison of diversification methods

6.1 Experiment setup

We evaluate the effectiveness of the proposed framework over two standard collections used for the diversity task in the TREC Web track (Clarke et al. 2009a, 2010). The first collection is denoted as **TREC09**, which contains 50 queries and uses the ClueWeb09

Category B as the document collection. The second collection is denoted as **TREC10**, which contains 48 valid queries with judgment files and uses the ClueWeb09 Category A collection as the document collection. The average number of subtopics per query is 4.83 for TREC09 and 4.36 for TREC10. The preprocessing involves stemming with Porter stemmer, stop word removal and deleting spam documents from the collection (Cormack et al. 2010). The performance is measured with several official measures including α-*nDCG*, where α is set to be 0.5, and *ERR-IA* at two retrieval depths, top 10 and top 20 documents. α-*nDCG@10* is used as the primary measure.

The baseline systems include: (1) *NoDiversity* which ranks documents based on their original relevance scores computed using Dirichlet method (Zhai and Lafferty 2001) as retrieval function; (2) *MMR* which uses Maximal Marginal Relevance method (Carbonell and Goldstein 1998) to re-rank documents.

We have derived five diversification methods as shown in Table 1. As discussed in Sect. 4.3, *SUM* is similar to the *WUME* (Yin et al. 2009). *PCOV* is similar to *xQuAD* (Santos et al. 2010b)[1] and *IA-SELECT* (Agrawal et al. 2009). The main differences between these two methods and *PCOV* are that they rely on external resources to extract subtopics and use different methods to instantiate the components in the function. Therefore, *PCOV* and *SUM* can be regarded as two strong baselines because they are similar to the state-of-the-art techniques.

The proposed optimization framework assumes that the subtopics of a query have been identified. We conduct two sets of experiments. In the first set of experiments, we use subtopics extracted from the collection to test the performances of the methods in real diversification systems. However, the effectiveness of the subtopic extraction method may affect the performance comparison. We therefore use the real query subtopics for diversification in the second set of experiments.

6.2 Performance comparison with extracted subtopics

We now report the performance when we extract query subtopics from the collection. The existing diversification methods use the topic mining method (Carterette and Chandar 2009) to extract query subtopics from the collection or external resources to extract subtopics (Santos et al. 2010b). However, the topic mining method is very time-consuming. External resources, i.e., query suggestions from search engines, are independent of the collection and the subtopics may not represent the relevant information of the query in the collection. Therefore, we use a pattern-based method (Zheng et al. 2011) to extract subtopics of the query. It extracts each group of terms that frequently co-occur in the retrieved documents of the query as a subtopic candidate of the query. It then computes the semantic similarity between the subtopic candidate and the query based on the average mutual information between the subtopic terms and query terms. The top-ranked subtopic candidates are selected as the subtopics of the query.

We first compare the performance of different diversification methods when using Dirichlet method (Zhai and Lafferty 2001) as retrieval function. The parameter μ is set to be 500. Table 3 shows the optimal performance of different diversification based on α-*nDCG* at different retrieval depths. The Wilcoxon signed rank tests compare the new functions with the existing methods, i.e., *NoDiversity, MMR, SUM* and *PCOV*. Table 4 shows the parameter values corresponding to the performances in Table 3. *SUPP* is the

[1] In fact, a variant of *xQuAD* achieves the best TREC performance (Clarke et al. 2009; McCreadie et al. 2009).

Table 3 Optimal performances when using extracted subtopics. +, *, ▲ and ◆ means improvement over *NoDiversity*, *MMR*, *SUM* and *PCOV*, respectively, are statistically significant ($p = 0.05$ in Wilcoxon test)

Collections	Methods	*α-nDCG@10*	*α-nDCG@20*	*ERR-IA@10*	*ERR-IA@20*
TREC09	*NoDiversity*	0.2122	0.2462	0.1414	0.1494
	MMR	0.2168	0.2538	0.1436	0.1523
	SUM	0.2511	0.2758	0.1749	0.1809
	PCOV	0.2540	0.2768	0.1764	0.1822
	LOG	0.2520+	0.2767+	0.1758+	0.1819 +
	EVAL	0.2510+	0.2771	0.1759	0.1813
	SQR	**0.2693**+*▲ ◆	**0.2907**+, ▲, ◆	**0.1886**+*	**0.1943**+
TREC10	*NoDiversity*	0.2269	0.2634	0.1735	0.1833
	MMR	0.2339	0.2746	0.1756	0.1868
	SUM	0.2858	0.3034	0.2253	0.2304
	PCOV	0.2924	0.3084	0.2305	0.2353
	LOG	0.2884+*	0.3058+*	0.2282+*	0.2334+*
	EVAL	0.2903+*▲	0.3063+*▲	0.2287+*▲	0.2334+*▲
	SQR	**0.2939**+*	**0.3252**+*	**0.2357**+*	**0.2447**+*

Bold values are best performances in different measures

Table 4 Optimal values of parameters when using extracted subtopics

Collections	Methods	*SUPP*	*SUB*	*TERM*	DOC	λ
TREC09	*SUM*	4	5	30	20	0.6
	PCOV	4	5	10	20	0.6
	LOG	3	3	20	20	0.7
	EVAL	20	5	30	20	0.6
	SQR	3	4	10	20	0.3
TREC10	*SUM*	2	3	15	20	0.6
	PCOV	2	3	15	20	0.6
	LOG	2	3	15	20	0.6
	EVAL	2	3	20	20	0.6
	SQR	2	6	20	70	0.3

minimum number of documents that each subtopic candidate must appear (Zheng et al. 2011), *SUB* is the number of extracted subtopics, *TERM* is the number of terms to use in each subtopic, *DOC* is the number of re-ranked documents using diversification functions and λ is the parameter of functions in Table 1. *EVAL* has another parameter α in Equation (25) whose optimal values are 0.4 on TREC09 and 0.6 on TREC10. These parameter values are the best possible values in each method. We have the following observations:

- All the subtopic-based diversification methods are effective to diversify search results, and they are more effective than *MMR*. Similar observations were reported in studies of Santos et al. (2010b).
- *SQR* outperforms all other diversification methods. This provides empirical supports for the analysis in Sect. 4 The significance tests show that the optimal performance of *SQR* is significantly better, i.e., with p-value smaller than 0.05, than the existing

methods based on α-*nDCG@10* on TREC09 collections. However, it cannot significantly outperform the existing methods on TREC10 collections. We will analyze these methods in detail in the rest of this section.

– *SUM* often performs worse than the other methods. This is expected because *SUM* does not consider the subtopic coverage of previously selected documents. Again, this confirms our analytical results in the previous section.

We also compare our results with the top runs of TREC here. The performance of *SQR* is better than the 4*th* best run of TREC09 whose α-*nDCG@10* value is 0.250 on TREC09 collection and is close to the 6*th* best run of TREC10 whose *ERR-IA@20* value is 0.248. However, they are not directly comparable because we use different baseline functions to retrieve and use different parameter tuning processes. Another observation is that *PCOV* result in this paper is worse than *xQuAD*, whose α-*nDCG@10* value is 0.282, on TREC09 collection, although they have similar diversification function. This is because that they use different baseline functions, component estimation methods and subtopic extraction methods.

The results in Table 3 are the best performances of each method. We then test the robustness of these methods. Table 5 shows the 5-fold cross-validation results based on $\alpha - nDCG@10$ on TREC09 and TREC10 collections. It tunes the values of all parameters shown in Table 4 and the parameter α in *EVAL*. It shows that *SQR* is more robust than existing functions, i.e., *SUM* and *PCOV*, on both collections. What's more, *LOG* and *EVAL* also outperform the existing functions on TREC09 and TREC10 collection, respectively.

The diversification methods have different diversification properties as we analyzed in Sect. 5 The queries also have different diversification features. For example, some queries have more subtopics while other queries have less subtopics. It is interesting to see the effect of different methods in different kinds of queries. We compare the average performance of each method in queries with different number of real subtopics and report the results in Table 6. Queries are divided into 5 bins according to the number of subtopics. An interesting observation is that when the number of subtopics becomes large, i.e., the relevant documents are more diverse, *SQR* performs the best among all the diversification methods. However, *SQR* does not perform best when the number of subtopics is 3 or 5. This indicates there is some potential to combine different diversification methods based on the number of subtopics and we leave this as future work.

Table 5 Cross-validation results of the diversification methods on α-*nDCG@*10 over all parameters when using Dirichlet and extracted subtopics

Collections	Methods	Train		Test	
		Average	Deviation	Average	Deviation
TREC09	*SUM*	0.2529	0.0144	0.2266	0.0682
	PCOV	0.2570	0.0161	0.2296	0.0599
	LOG	0.2526	0.0149	0.2309	0.0637
	EVAL	0.2525	0.0114	0.2294	0.0539
	SQR	0.2697	0.0137	**0.2627**	0.0587
TREC10	*SUM*	0.2867	0.0196	0.2792	0.0795
	PCOV	0.2947	0.0213	0.2731	0.0977
	LOG	0.2899	0.0213	0.2738	0.0986
	EVAL	0.2910	0.0197	**0.2842**	0.0830
	SQR	0.2948	0.0202	0.2829	0.0871

Bold values are best performances in test results

Table 6 Performance comparison (*α-nDCG@10*) on queries with different number of subtopics

Collections	Methods	Number of groundtruth subtopics				
		3	4	5	6	≥ 7
TREC09	*SUM*	**0.4691**	0.2562	0.2386	0.2061	0.0671
	PCOV	0.4632	0.2586	0.2428	0.2133	0.070
	LOG	0.4075	0.2605	0.2443	0.2308	0.0671
	EVAL	0.4496	0.2437	**0.2586**	0.2216	0.0383
	SQR	0.4434	**0.2901**	0.2442	**0.2435**	**0.0765**
	Number of queries	6	16	13	11	4
TREC10	*SUM*	0.3435	0.2566	0.3465	0.2252	0.0538
	PCOV	**0.3575**	0.2573	**0.3622**	0.2206	0.0856
	LOG	0.3464	0.2580	0.3523	0.2266	0.0538
	EVAL	0.3495	0.2582	0.3539	0.2271	0.0867
	SQR	0.3358	**0.2692**	0.3178	**0.2838**	**0.1173**
	Number of queries	11	18	10	8	1

Bold values are best performances in different categories of queries

Table 7 Optimal performance of the diversification methods using different traditional retrieval functions based on *α-nDCG@10* (extracted subtopics). +, ▲ and ◆ means improvement over *NoDiversity*, *SUM* and *PCOV*, respectively, are statistically significant ($p = 0.05$ in Wilcoxon test)

Collection	Methods	Pivot	Okapi	Dirichlet	F2exp
TREC09	*NoDiversity*	0.1530	0.2495	0.2122	0.2351
	SUM	0.2571	0.2642	0.2511	0.2609
	PCOV	0.2633	0.2593	0.2540	0.2448
	LOG	0.2608+	0.2627+	0.2520+	0.2599+
	EVAL	0.2651+	0.2585+	0.2510+	0.2607+
	SQR	**0.2688**+	**0.2720**+	**0.2693**+▲◆	**0.2779**+◆
TREC10	*NoDiversity*	0.2084	0.2252	0.2269	0.2240
	SUM	0.2721	0.2905	0.2858	0.2938
	PCOV	0.2761	0.3065	0.2924	0.2992
	LOG	0.2767	0.2892	0.2884+	0.2938
	EVAL	0.2786	0.2905	0.2903+▲	0.2938
	SQR	**0.2828**	**0.3213**+▲	**0.2939**+	**0.3163**+

Bold values are best performances in different measures

As discussed earlier, one advantage of the derived diversification methods is that they can be combined with any retrieval functions. In order to evaluate the effectiveness of these diversification methods for different retrieval functions, we use four state-of-the-art retrieval functions and report the optimal performance of five diversification methods for these four retrieval models on both collections in Table 7. These retrieval functions include pivoted normalization method (Singhal et al. 1996), Dirichlet (Zhai and Lafferty 2001), axiomatic retrieval function, i.e., F2exp (Fang and Zhai 2005), and Okapi (Robertson and Walker 1999) which was also applied in *xQuAD* (Santos et al. 2010b). When using one retrieval function, we use the same retrieval function to compute components in

Table 8 Optimal performances when using real subtopics. +, ▲ and ◆ means improvement over *NoDiversity*, *SUM* and *PCOV*, respectively, are statistically significant ($p = 0.05$ in Wilcoxon test)

Collections	Methods	*α-nDCG@10*	*α-nDCG@20*	*ERR-IA@10*	*ERR-IA@20*
TREC09	*NoDiversity*	0.2122	0.2462	0.1414	0.1494
	SUM	0.2731	0.2915	0.1901	0.1952
	PCOV	0.2813	0.2927	0.1902	0.1948
	LOG	0.2736+	0.2911+	0.1896+	0.1947+
	EVAL	0.2740+	0.2915+	0.1902+	0.1953+
	SQR	**0.2819**+	**0.2981**+	**0.1970**+	**0.2013**+
TREC10	*NoDiversity*	0.2269	0.2634	0.1735	0.1833
	SUM	0.3643	0.3761	0.2813	0.2858
	PCOV	0.3648	0.3785	0.2846	0.2894
	LOG	0.3643+	0.3769+	0.2826+	0.2873+
	EVAL	0.3643+	0.3761+	0.2813+	0.2858+
	SQR	**0.3681**+	**0.3799**+	**0.2852**+	**0.2898**+

Bold values are best performances in different measures

Table 9 Cross-validation results of the diversification methods on *α-nDCG@10* over all parameters when using Dirichlet and real subtopics

Collections	Methods	Train		Test	
		Average	Deviation	Average	Deviation
TREC09	*SUM*	0.2752	0.0199	0.2644	0.0796
	PCOV	0.2831	0.0179	0.2649	0.0826
	LOG	0.2754	0.0201	0.2635	0.0802
	EVAL	0.2758	0.0201	0.2629	0.0811
	SQR	0.2843	0.0207	**0.2660**	0.0814
TREC10	*SUM*	0.3652	0.0171	0.3564	0.0745
	PCOV	0.3656	0.0176	0.3584	0.0750
	LOG	0.3643	0.0171	0.3643	0.0733
	EVAL	0.3652	0.0171	0.3549	0.0734
	SQR	0.3683	0.0165	**0.3658**	0.0708

Bold values are best performances in test results

diversification functions and normalize these scores as the probability in the functions of Table 1. We tuned the parameters in these retrieval functions. The parameters values are: (1) the values of s are 0.1 on TREC09 collection and 0.2 on TREC10 collection in Pivot; (2) k_1, b and k_3 are set to be 1.2, 0.75 and 1000, respectively in Okapi; (3) μ is set to be 500 in Dirichlet and (4) the values of s are set to be 0.8 on TREC09 and 0.9 on TREC10, and k is set to be 0.35 in F2exp. We can see that *SQR* is the most robust diversity function and can perform best when using any traditional retrieval models. Note that the diversity performance is closely related to the retrieval function. If we use a stronger baseline that considers factors other than keyword matching, such as the methods used by top runs by TREC participants (Clarke et al. 2009a, 2010), the diversity performance would be expected to increase.

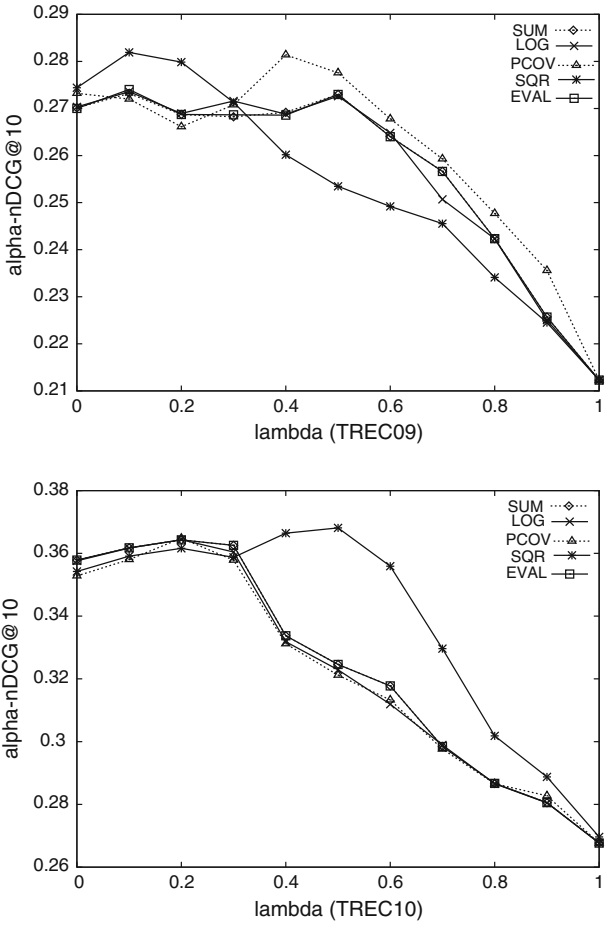

Fig. 1 The influence of λ on the TREC09 (*top*) and TREC10 (*bottom*) collections

6.3 Performance comparison using real subtopics

We also compare different diversification methods using real subtopics. This would allow us to factor out the effects of subtopic quality and directly compare the effectiveness of different diversification methods.

Table 8 shows the optimal performance of all the five diversification methods when we use real subtopics from judgment file as the subtopics of the query and use Dirichlet method (Zhai and Lafferty 2001) to retrieve documents. We can see that the performance of these different diversification methods are similar, and *SQR* performs slightly better than the other four diversification methods. Table 9 shows the cross-validation results of these methods when tuning parameters λ and α. We can see that *SQR* consistently outperform the existing functions, i.e., *SUM* and *PCOV*, on both collection. *LOG* can also outperform existing functions on TREC10 collection.

We also examine the performance sensitivity with respect to parameter values of the diversification functions. All of the derived diversification methods as shown in Table 1 have the same parameter, λ, which controls the balance between relevance and diversity

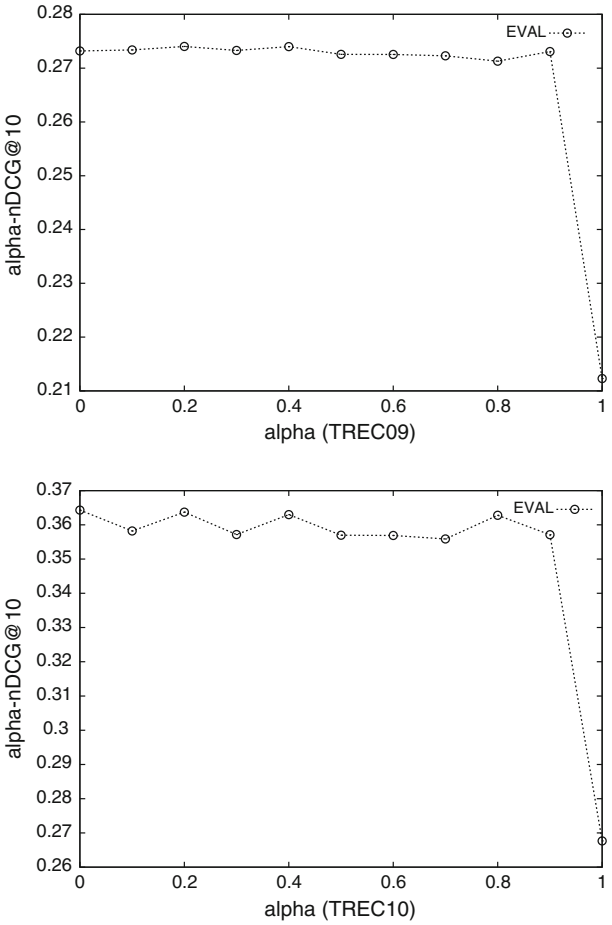

Fig. 2 The influence of α on the TREC09 (*top*) and TREC10 (*bottom*) collections

scores. In addition, *EVAL* has one more parameter α, which controls the impact of the individual documents in the coverage function.

Figure 1 shows the performance sensitivity curve for λ. When λ is 1, all the methods have the same performance since the documents are selected based on only relevance scores. We can also see that, when λ is 0, the performance is much better than the performance when λ is 1. The reason is that when λ is 0, the objective function is only related to the diversity score which is computed based on the relevance score between documents and query subtopics. Since the subtopics are chosen from judgment file, they have real good quality, which would lead to better performance. We can imagine that, when the quality of query subtopics decreases, the performance when λ is 0 would be worse than the one shown on the plot. Moreover, it is interesting to see that the lines of *SQR* and *PCOV* have different trends with other methods. However, the performances of *SUM* and *EVAL* in Fig. 1 are very close to each other when we use real subtopics. The results in Table 9 also shows that *EVAL* does not work well. These results indicates that the component $(1-\alpha)^{\sum_{d' \in D} cov(s,d')}$ in (25) is not effective in diversification.

Figure 2 shows the results of diversity function *EVAL* with different values of α. We can see that *EVAL* is not sensitive to the parameter when α is smaller than 1 and the

performance change is small when using different values. The result of *EVAL* is the same as *NoDiversity* when α is 1.

7 Conclusions

The task of search result diversification is to return a list of documents that are not only relevant to a query but also diverse to cover multiple subtopics of the query. In this paper, we propose to study this problem in a coverage-based optimization framework based on explicit query subtopic modeling. In this framework, we propose to model diversity scores based on the coverage of query subtopics and then discuss three strategies to define coverage functions. Not every function can be coverage function. A coverage function needs to be a submodular function so that we can use a greedy algorithm to iteratively select documents for diversity. Each coverage function corresponds to a diversification method. We derive five diversification methods in this paper, and show that the obtained methods include not only several existing methods, but also new ones which have not been studied before. One of the method, i.e., *SQR*, not only has the desired *favoring diversity* and *novelty emphasis* properties that the existing methods do not have, but also can consistently outperform the existing methods in the experiments.

Our work opens up many interesting future research directions. First, we plan to define more coverage functions in our framework and thus derive effective retrieval functions. For example, we can use better evaluation measures (Sakai and Song 2011) to derive diversification functions with desired properties. We can also combine different types of coverage functions (e.g., linear combination) to define more sophisticated ones in our framework. Second, it would be interesting to derive reasonable coverage functions based on a set of desirable properties such as the three we discuss in this paper.

Acknowledgments This material is based upon work supported by the National Science Foundation under Grant Number IIS-1017026. We thank the reviewers for their useful comments.

Appendix

Theorem 1 *$div_{SUM}(q, D)$ is a submodular function with respect to D.*

Proof In order to prove that $div_{SUM}(q, D)$ is submodular, we need to prove that 12 holds according to the analysis in Sect. 4.1 Based on 16,

$$DIFF_{SUM}(s,d,A) - DIFF_{SUM}(s,d,B) = cov(s,d) - cov(s,d) = 0.$$

Therefore, $div_{SUM}(q, D)$ is a submodular function.

Theorem 2 *$div_{LOG}(q, D)$ is a submodular function with respect to D.*

Proof Based on (18),

$$\begin{aligned} &DIFF_{LOG}(s,d,A) - DIFF_{LOG}(s,d,B) \\ &= \log\left(1 + \frac{cov(s,d)}{1 + \sum_{d' \in A} cov(s,d')}\right) - \log\left(1 + \frac{cov(s,d)}{1 + \sum_{d' \in B} cov(s,d')}\right). \end{aligned}$$

It is clear that $\sum_{d' \in A} cov(s,d') \leq \sum_{d' \in B} cov(s,d')$. Therefore,

$$\log\left(1+\frac{cov(s,d)}{1+\sum_{d'\in A} cov(s,d')}\right) \geq \log\left(1+\frac{cov(s,d)}{1+\sum_{d'\in B} cov(s,d')}\right),$$

and $DIFF_{SUM}(s, d, A) - DIFF_{SUM}(s, d, B) \geq 0$. $cov_{LOG}(s, D)$ is a submodular function.

Theorem 3 *$div_{PCOV}(q, D)$ is a submodular function with respect to D*

Proof Based on (20),

$$\begin{aligned}
&DIFF_{PCOV}(s,d,A) - DIFF_{PCOV}(s,d,B)\\
&= cov(s,d)\cdot\prod_{d'\in A}(1-cov(s,d')) - cov(s,d)\cdot\prod_{d'\in B}(1-cov(s,d'))\\
&= cov(s,d)\cdot\left(1-\prod_{d'\in B\setminus A}(1-cov(s,d'))\right)\cdot\prod_{d'\in A}(1-cov(s,d'))\\
&\geq 0.
\end{aligned}$$

Thus, $cov_{PCOV}(s, D)$ is a submodular function.

Theorem 4 *$div_{SQR}(q, D)$ is a submodular function with respect to D*

Proof Based on (23),

$$\begin{aligned}
&DIFF_{PCOV}(s,d,A) - DIFF_{PCOV}(s,d,B)\\
&= cov(s,d)\cdot 2\cdot\sum_{d'\in B\setminus A}(cov(s,d')) \geq 0.
\end{aligned}$$

Thus, $cov_{SQR}(s, D)$ is a submodular function.

Theorem 5 *$div_{EVAL1}(q,D)$ is not a submodular function with respect to D*

Proof Based on (24),

$$\begin{aligned}
&DIFF_{EVAL1}(s,d,A) - DIFF_{EVAL1}(s,d,B)\\
&= \frac{1}{|A|+1}\cdot\left(cov(s,d)-\frac{\sum_{d'\in A} cov(s,d')}{|A|}\right)\\
&-\frac{1}{|B|+1}\cdot\left(cov(s,d)-\frac{\sum_{d'\in B} cov(s,d')}{|B|}\right)\\
&= cov(s,d)\cdot\left(\frac{1}{|A|+1}-\frac{1}{|B|+1}\right)+\frac{\sum_{d'\in B\setminus A}(cov(s,d')}{|B|\cdot(|B|+1)}\\
&-\left(\frac{1}{|A|\cdot(|A|+1)}-\frac{1}{|B|\cdot(|B|+1)}\right)\cdot\sum_{d'\in A} cov(s,d'),
\end{aligned}$$

and this may be smaller than 0. Thus, $cov_{EVAL1}(s,D)$ is not a submodular function.

Theorem 6 *$div_{EVAL2}(q,D)$ is a submodular function with respect to D*

Proof Based on (24),

$$\begin{aligned}
&DIFF_{EVAL2}(s,d,A) - DIFF_{EVAL2}(s,d,B)\\
&= cov(s,d)\cdot\left(\frac{\prod_{d'\in A}(1-cov(s,d'))}{|A|+1}-\frac{\prod_{d'\in B}(1-cov(s,d'))}{|B|+1}\right)
\end{aligned}$$

We know that $\prod_{d' \in A}(1 - cov(s, d')) \geq \prod_{d' \in B}(1 - cov(s, d'))$ since

$$\begin{aligned}
&\prod_{d' \in A}(1 - cov(s, d')) - \prod_{d' \in B}(1 - cov(s, d')) \\
&= \prod_{d' \in A}(1 - cov(s, d')) \cdot (1 - \prod_{d' \in B \setminus A}(1 - cov(s, d'))) \\
&\geq 0,
\end{aligned}$$

and $|A| \leq |B|$. Therefore $DIFF_{EVAL2}(s, d, A) - DIFF_{EVAL2}(s, d, B) \geq 0$. Thus, $cov_{EVAL2}(s,D)$ is a submodular function.

Theorem 7 *$div_{EVAL3}(q,D)$ is a submodular function with respect to D*

Proof Based on (25),

$$\begin{aligned}
&DIFF_{EVAL3}(s, d, A) - DIFF_{EVAL3}(s, d, B) \\
&= cov(s, d) \cdot \left((1 - \alpha)^{\sum_{d' \in A} cov(s, d')} - (1 - \alpha)^{\sum_{d' \in B} cov(s, d')} \right)
\end{aligned}$$

We can find that $(1 - \alpha)^{\sum_{d' \in A} cov(s, d')} \geq (1 - \alpha)^{\sum_{d' \in B} cov(s, d')}$ since $\sum_{d' \in A} cov(s, d') \leq \sum_{d' \in B} cov(s, d')$. Therefore, *EVAL3* is a submodular function since $DIFF_{EVAL3}(s, d, A) - DIFF_{EVAL3}(s, d, B) \geq 0$. Thus, $cov_{EVAL3}(s,D)$ is a submodular function.

References

Agrawal, R., Gollapudi, S., Halverson, A., & Ieong, S. (2009). Diversifying search results. In *Proceedings of WSDM'09*.

Boyce, B. (1982). Beyond topicality: A two stage view of relevance and the retrieval process. *Information Processing and Management, 18*(3), 105–109

Carbonell, J., & Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of SIGIR'98*, pp. 335–336.

Carterette, B., & Chandar, P. (2009). Probabilistic models of novel document rankings for faceted topic retrieval. In *Proceedings of CIKM'09*.

Chapelle, O., Metzler, D., Zhang, Y., & Grinspan, P. (2009). Expected reciprocal rank for graded relevance. In *Proceedings of CIKM'09*.

Chen, H., & Karger, D. R. (2006). Less is more: Probabilistic models for retrieving fewer relevant documents. In *Proceedings of SIGIR'06*.

Clarke, C. L. A., Craswell, N., & Soboroff, I. (2009). Overview of the trec 2009 web track. In *Proceedings of TREC'09*.

Clarke, C. L. A., Kolla, M., & Vechtomova, O. (2009). An effectiveness measure for ambiguous and underspecified queries. In *Proceedings of ICTIR'09*.

Clarke, C. L. A., Craswell, N., Soboroff, I., & Cormack, G. V. (2010). Preliminary overview of the trec 2010 web track. In *Proceedings of TREC'10*.

Clarke, C. L. A., Kolla, M., Cormack, G. V., Vechtomova, O., Ashkan, A., Buttcher, S. et al. (2008). Novelty and diversity in information retrieval evaluation. In *Proceedings of SIGIR'08*.

Cormack G. V., Smucker, M. D., & Clarke, C. L. A. (2010). Efficient and effective spam filtering and re-ranking for large web datasets. In http://arxiv.org/abs/1004.516

Craswell, N., Fetterly, D., Najork, M., Robertson, S., & Yilmaz, E. (2009). Microsoft research at trec 2009. In *Proceedings of TREC'09*.

Demidova, E., Fankhauser, P., Zhou, X., & Nejdl, W. (2010). Divq: Diversification for keyword search over structured databases. In *Proceedings of SIGIR'10*.

Fang, H., & , Zai, C. (2005). An exploration of axiomatic approaches to information retrieval. In *Proceedings of SIGIR'05*.

Goffman, W. (1964). A search procedure for information retrieval. *Information Storage and Retrieval, 2*, 73–78.

Gollapudi, S., & Sharma, A. (2009). An axiomatic approach for result diversification. In *Proceedings of WWW'09*.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The element of statistical learning: Data mining, inference and prediction*. Berlin: Springer.

Khuller, S., Moss, A., & NaorJure, J. (1999). The generalized maximum coverage problem. *Information Processing Letters, 70*(1), 39–45.

Lafferty, J., & Zhai, C. (2001). Document language models, query models, and risk minimization for information retrieval. In *Proceedings of SIGIR'01*, Sept 2001.

Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., & Glance, N. (2007). Cost-effective outbreak detection in networks. In *Proceedings of KDD'07*, 2007.

Macdonald, C., Wang, J., & Clarke, C. (2011). Ecir2011 ddr workshop proceedings. In *Proceedings of DDR'11*, 2011.

McCreadie, R., Macdonald, C., Ounis, I., Peng, J., & Santos, R. (2009). University of glasgow at trec 2009: Experiments with terrier. In *Proceedings of TREC'09*.

Nemhauser, G. L., Wolsey, L. A., & Fisher, M. L. (1978). An analysis of approximations for maximizing submodular set functions. *Mathematical Programming, 14*(1), 265–294.

Ponte, J., & Croft, W. B. (1998). A language modeling approach to information retrieval. In *Proceedings of the ACM SIGIR'98*.

Radlinsk, F., & Dumais, S. T. (2006). Improving personalized web search using result diversification. In *Proceedings of SIGIR'06*.

Radlinski, F., Bennett, P. N., Carterette, B., & Joachims, T. (2009). Redundancy, diversity and interdependent document relevance. In *Proceedings of the IDR'09 Workshop*.

Robertson, S. E., & Walker, S. (1999). Okapi/keenbow at TREC-8. In E. M. Voorhees & D. K. Harman (Eds.), *The eighth text REtrieval conference (TREC 8)*. NIST Special Publication 500-246.

Sakai, T., & Song, R. (2011). Evaluating diversified search results using per-intent graded relevance. In *Proceedings of SIGIR'11*.

Santos, R. L., & Ounis, I. (2011). Diversifying for multiple information needs. In *Proceedings of DDR'11*.

Santos, R. L. T., Macdonald, C., & Ounis, I. (2010). Selectively diversifying web search results. In *Proceedings of CIKM'10*.

Santos, R. L. T., Macdonald, C., & Ounis, I. (2010). Exploiting query reformulations for web search result diversification. In *Proceedings of WWW'10*.

Santos, R. L. T., Peng, J., Macdonald, C., & Ounis, I. (2010). Explicit search result diversification through sub-queries. In *Proceedings of ECIR'10*.

Singhal, A., Buckley, C., & Mitra, M. (1996). Pivoted document length normalization. In *Proceedings of the 1996 ACM SIGIR conference on research and development in information retrieval*, pp. 21–29.

White, R. W., & Roth, R. A. (2009) *Exploratory search: Beyond the query-response paradigm*. San Rafael, CA: Morgan and Claypool.

Yin, D., Xue, Z., Qi, X., & Davison, B. D. (2009). Diversifying search results with popular subtopics. In *Proceedings of TREC'09*.

Yue, Y., & Joachims, T. (2008). Predicting diverse subsets using structural svms. In *Proceedings of ICML'08*.

Zhai, C., Cohen, W., & Lafferty, J. (2003). Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. In *Proceedings of SIGIR'03*.

Zhai, C., & Lafferty, J. (2001). A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of SIGIR'01*.

Zheng, W., & Fang, H. (2010). University of delaware at diverstiy task of web track 2010. In *Proceedings of TREC'10*.

Zheng, W., & Fang, H. (2011). A comparative study of search result diversification methods. In *Proceedings of DDR'11*.

Zheng, W., Wang, X., Fang, H., & Cheng, H. (2011). An exploration of pattern-based subtopic modeling for search result diversification. In *Proceedings of JCDL'11*.