Inf Retrieval (2013) 16:452–483
DOI 10.1007/s10791-012-9215-y

SEARCH INTENTS AND DIVERSIFICATION

Mining subtopics from different aspects for diversifying search results

Chieh-Jen Wang · Yung-Wei Lin · Ming-Feng Tsai · Hsin-Hsi Chen

Received: 19 May 2012 / Accepted: 30 November 2012 / Published online: 18 December 2012
© Springer Science+Business Media New York 2012

Abstract User queries to the Web tend to have more than one interpretation due to their ambiguity and other characteristics. How to diversify the ranking results to meet users' various potential information needs has attracted considerable attention recently. This paper is aimed at mining the subtopics of a query either indirectly from the returned results of retrieval systems or directly from the query itself to diversify the search results. For the indirect subtopic mining approach, clustering the retrieval results and summarizing the content of clusters is investigated. In addition, labeling topic categories and concept tags on each returned document is explored. For the direct subtopic mining approach, several external resources, such as Wikipedia, Open Directory Project, search query logs, and the related search services of search engines, are consulted. Furthermore, we propose a diversified retrieval model to rank documents with respect to the mined subtopics for balancing relevance and diversity. Experiments are conducted on the ClueWeb09 dataset with the topics of the TREC09 and TREC10 Web Track diversity tasks. Experimental results show that the proposed subtopic-based diversification algorithm significantly outperforms the state-of-the-art models in the TREC09 and TREC10 Web Track diversity tasks. The best performance our proposed algorithm achieves is α-nDCG@5 0.307, IA-P@5 0.121, and α#-nDCG@5 0.214 on the TREC09, as well as α-nDCG@10 0.421, IA-P@10 0.201, and α#-nDCG@10 0.311 on the TREC10. The results conclude that the subtopic mining technique with the up-to-date users' search query logs is the most

C.-J. Wang · Y.-W. Lin · H.-H. Chen (✉)
Department of Computer Science and Information Engineering, National Taiwan University, No. 1, Sec. 4, Roosevelt Rd., Taipei 10617, Taiwan
e-mail: hhchen@ntu.edu.tw

C.-J. Wang
e-mail: cjwang@nlg.csie.ntu.edu.tw

Y.-W. Lin
e-mail: ywlin@nlg.csie.ntu.edu.tw

M.-F. Tsai
Department of Computer Science and Program in Digital Content and Technologies, National Chengchi University, No. 64, ZhiNan Rd Sec. 2, Taipei 11605, Taiwan
e-mail: mftsai@nccu.edu.tw

effective way to generate the subtopics of a query, and the proposed subtopic-based diversification algorithm can select the documents covering various subtopics.

Keywords Diversified retrieval · Subtopic mining · Search result re-ranking

1 Introduction

Users' queries represented by a few keywords usually contain a certain extent of ambiguity (Spärck-Jones et al. 2007). An ambiguous query may refer to multiple aspects or have more than one interpretation. In this paper, different interpretations and aspects associated with a query are called *subtopics*. For example, the query "dinosaur" may refer to three subtopics: (1) a paleontological science, (2) a multimedia recreation, and (3) a scenic spot. Each subtopic may also contain several sub–sub-topics, for example, "comics", "television shows", "films", and "music bands" with respect to the subtopic of "a multimedia recreation". In addition, even a query with a clearly faceted interpretation might still be underspecified, because it is not clear which subtopic of the interpretation is actually desirable for users. For instance, the faceted query "air travel information" may contain different subtopics, such as (1) information on air travel, airports, and airlines, (2) restrictions for checked baggage during air travel, and (3) websites that collect statistics and report about airports.

Traditional information retrieval models, such as the Boolean model and the vector space model typically only consider the relevance between a query and documents. These retrieval models treat every input query as a clear, well-defined representation and completely neglect any sort of ambiguity. This negligence results in the top ranked documents possibly containing too much relevant information on the same subtopic. This might increase users' search time for distinguishing whether the retrieved documents contain redundant information. In addition, some retrieval models anticipate that the underlying meaning of a submitted query is always the most popular subtopic. These models may focus the retrieval process on popular/particular subtopics too much. This postulation may have risks with a wrong guess (e.g., the user information need is different from the most popular subtopic), which could leave users unsatisfied. For maximizing the satisfaction of different users, a retrieval model has to select a list of documents that are not only relevant to some popular subtopics, but also covers different subtopics. Users can quickly find relevant information that may be interesting if search results are diversified. Nevertheless, how to balance the relevance and the diversity of search results is a trade-off. On the one hand, too many subtopics may provide diversified information but introduce many irrelevant documents, which cause a relevance issue. On the other hand, if only the similarity between a query and documents is considered, too many documents belonging to the same subtopics are retrieved, which causes a diversity issue.

In recent years, using subtopics of a query for diversifying the retrieved documents has received considerable attention (Song et al. 2011). The broad topic associated with an ambiguous or unclear query can be decomposed into a set of subtopics. This provides an opportunity to deal with the problem of search result diversification, as we can employ the clues from the subtopics to rank a diverse ranking list based on optimizing the maximum coverage of the subtopics. In this paper, we introduce a novel framework for search result diversification that exploits the subtopics embedded in queries and ranks the retrieved documents based on these discovered subtopics. Several methods are proposed for mining subtopics from different aspects, such as the retrieved documents, the search query logs, and the related search services provided by the commercial search engines.

Theoretically, the diversified retrieval models should provide a ranking list of documents that has the maximum coverage and minimum redundancy with respect to the possible subtopics underlying a query. Moreover, the covered subtopics should also reflect their relative importance for the query, as perceived from most users (Yin et al. 2009; Agrawal et al. 2009). For example, a query "java" may have three subtopics (a programming language, coffee, and an island), where the subtopic of island attracts less interest than the subtopic of programming language. The subtopic of programming language should be relatively more important than the subtopic of island for the query "java".

In this paper, we propose a subtopic-based diversified retrieval framework that first uncovers different subtopics embedded in a query, then assigns a weight for each mined subtopic to describe its importance, and finally estimates the relevance of the retrieved documents to each mined subtopic for diversifying search results. The proposed framework not only keeps the quality of relevance, but also re-ranks the top-ranked retrieved results to cover multiple important subtopics. Specifically, there are three components in the proposed diversification framework, the *richness* of subtopics, the *importance* of subtopics, and the *novelty* of subtopics. The *richness* part aims at measuring how many subtopics are covered by a document, the *importance* part estimates the importance of the subtopics of a query, and the *novelty* part computes how many subtopics have already been covered by the previously retrieved documents. With these three aspects, the proposed document ranking algorithm uses a greedy-like strategy to select a list of documents that can cover as many multiple and relatively important subtopics as possible.

We conduct a series of experiments to evaluate the effectiveness of the proposed subtopic mining techniques and the diversification ranking algorithms. The experimental datasets are the ClueWeb09 Category A and Category B test collections with the topics of the TREC09 and TREC10 Web Track diversity tasks. The subtopic-based diversified retrieval framework has a large impact on the effectiveness for search result diversification, as shown in the experimental results. Compared with the state-of-the-art models in the TREC09 and TREC10, the proposed diversified retrieval framework significantly improves the diversity of search results, especially when integrating multiple aspects of resources.

The remainder of this paper is organized as follows. The related works are presented and compared in Sect. 2. Section 3 describes the subtopic mining methods and document diversification algorithms. In Sect. 4, the datasets used for experiments are described. The experimental results are reported and discussed in Sect. 5. We finally conclude our work and provide several directions for future work in Sect. 6.

2 Related works

In this section, we review some query understanding approaches to mining search intents and improving the performance of information retrieval systems. Next, we survey several previous studies about diversifying search results. Finally, we address the contributions of our subtopic-based diversified retrieval framework for search result diversification.

2.1 Query understanding

Many information retrieval models have benefited from taking into account the users' search intent. These models generally have relied on predefined categories to predict underlying search intents of queries for improving the search performance (Rose and Levinson 2004; Chang et al. 2006). Understanding users' search intents of queries can be

achieved using different types of external resources, e.g., Wikipedia, Open Directory Project (ODP) (Hu et al. 2009).

To realize the meanings of queries, several approaches have adopted taxonomies to classify queries into predefined search intent categories of different granulations. Broder (2002) divided query intent into navigational, informational, and transactional types. Nguyen and Kan (2007) characterized queries along four general facets of ambiguity, authority, temporal sensitivity, and spatial sensitivity. Manshadi and Li (2009) constructed a hybrid, generative grammar model based on probabilistic context-free rules for classifying queries into finer categories. Geng et al. (2008) applied the *k*-nearest neighbor (*k*-NN) classification algorithm to assign search intent categories of queries. Given an unseen query, a classifier generated by the *k*-NN classification algorithm was used to identify which training queries were similar to the unseen query before assigning the search intent of the most similar query to the unseen query. Hu et al. (2009) designed a random walk method using Wikipedia to predict query search intent. Understanding the search intent is helpful to improve search effectiveness. Radlinski and Joachims (2005) mined search intent from query chains and applied it to learning to rank algorithms. Boldi et al. (2008) employed the query-flow graphs to predict the search intent of queries for query recommendation.

2.2 Search result diversification

Diversifying search results has been studied and applied to different applications (Zhai et al. 2003; Radlinski et al. 2009; Santos et al. 2010b). Generally speaking, the previous works of search result diversification can be categorized into *implicit* or *explicit* approaches (Santos et al. 2010a). The implicit approaches assume the related documents will contain similar subtopics and can be regarded as redundant information. These similar documents might be demoted in the ranking list for diversification. The explicit approaches, in turn, directly model the subtopics of queries, and search the retrieved documents to maximize the coverage of the subtopics. The main difference between the two approaches is the implicit approaches identify the redundant information between a new document and the previous selected documents without considering the subtopics of a query explicitly.

For the implicit approaches, Carbonell and Goldstein (1998) proposed the maximal marginal relevance (MMR) to rank a retrieved document under a combination of a relevancy score with respect to a query and a dissimilarity score with respect to other similar documents selected at earlier ranks. Zhai et al. (2003) modeled relevance and redundancy based on the *KL*-divergence measure and a simple mixture model. Yue and Joachims (2008) maximized the word coverage to select the optimum set of diversified documents. The learned model selected documents for covering maximum distinct words with the greedy search. Chen and Karger (2006) presented a selection algorithm based on the Bayesian information retrieval framework for diversifying search results among the top ten previously visited results. Their selection algorithm estimated the documents based on the probability ranking principle and used pseudo-relevance feedback to search result diversification by negative feedback on the redundant documents. Vee et al. (2008) proposed several B+ tree based diversifying models to return a set of different answers on query answering diversification. Gollapudi and Sharma (2009) developed a set of natural axioms for diversification and utilized several diversified functions in their diversification framework. Wang and Zhu (2009) used mean–variance analysis to search result diversification based on the economic portfolio theory. Their algorithm estimated an uncertainty in terms of the "risk" in economic domain trade-off between the expected relevance of a set of

retrieved documents and the correlation between them, and it selected the right combination of relevant documents under the uncertainty estimation.

Different from the implicit approaches, subtopics can be modeled explicitly by queries for diversifying search results. Subtopics may be mined from a predefined taxonomy (Agrawal et al. 2009), related sub-queries, and suggested sub-queries (Santos et al. 2010a), etc. Radlinski and Dumais (2006) used the query–query reformulation records in search query logs to discover subtopics, and they diversified the retrieved results for improving effectiveness on personalized search. Agrawal et al. (2009) presented a systematic approach to diversifying search results through modeling subtopic from queries and documents by classification taxonomy for minimizing the risk of dissatisfaction of the average users. Carterette and Chandar (2009) identified subtopics by topic models and used a probabilistic ranking model to maximize the subtopic coverage rate for search result diversification. Santos et al. (2010a, b) uncovered subtopics by the query suggestion from search engines and proposed a probabilistic framework that estimated the diversity based on not only the documents' relevance to query subtopics, but also the relative importance associated with the query subtopics. Rafiei et al. (2010) identified subtopics by existing taxonomic information and regarded the problem of diversifying results as expectation maximization. They attempted to broaden the coverage of subtopics in the retrieved results. Welch et al. (2011) employed WordNet and Wikipedia to discover subtopics and presented a diversification algorithm especially suitable for informational queries where users may take more than one page to satisfy their needs.

Compared to the studies described above, the major contributions of this work are fourfold.

- Exploring various subtopic mining methods: A total of six subtopic mining methods are explored in this paper. These methods mine subtopics from different aspects, such as retrieved documents and external resources.
- Analyzing the effectiveness of the subtopic mining methods in depth: We analyze the effectiveness of the subtopics derived from different subtopic mining methods. A user study comparing the subtopics mined by different mining methods with the ground truth in the TREC09 and TREC10 is conducted.
- A novel subtopic-based diversified algorithm: The proposed diversification algorithm optimizes the estimation from three aspects, the *richness* of subtopics, the *importance* of subtopics, and the *novelty* of subtopics for diversifying search results.
- Thorough evaluation of the document ranking experiments: Through experiments on the ClueWeb09 Category A and Category B test collections with the topics of the TREC09 and TREC10, we demonstrate that our model significantly improves the performance of the state-of-the-art models proposed in the TREC09 and TREC10 Web Track diversity tasks.

3 A diversified retrieval system

Figure 1 shows our subtopic-based diversified retrieval framework, which contains two phases, subtopic mining and document ranking. The ClueWeb09 dataset is indexed by an information retrieval model, and the model reports a set of relevant documents after query submissions. In this study, the number of returned documents is set to one thousand. Several resources are employed to mine the subtopics of a query. The indirect mining methods mine subtopics from the retrieved documents for the given queries, and the direct

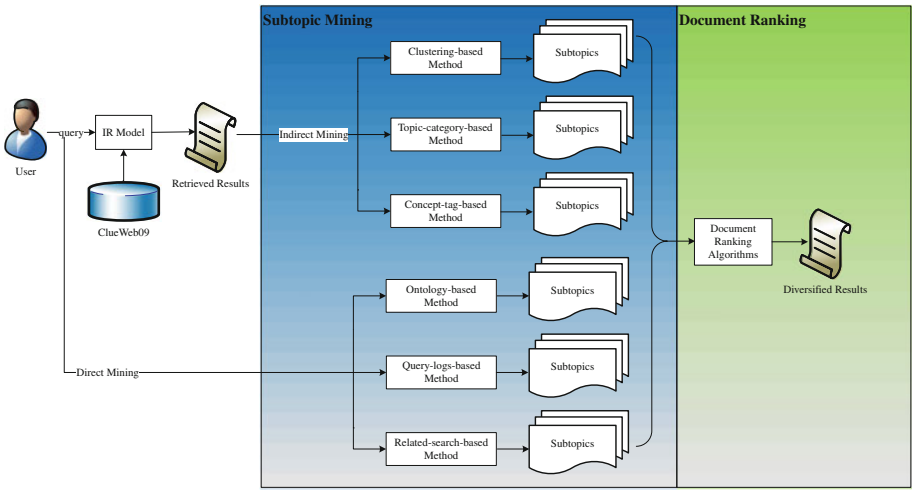

Fig. 1 A subtopic-based diversified retrieval framework

mining methods mine subtopics from the queries themselves. Three indirect subtopic mining methods, including a clustering-based method, a topic-category-based method, and a concept-tag-based method, are proposed. In addition, three direct subtopic mining methods, including an ontology-based method, a query-logs-based method, and a related-search-based method, are also introduced. The six mining methods are explored in the following sections. After the subtopic mining, we propose two document ranking algorithms based on the mined subtopics for search result diversification.

In the following sections, we begin by introducing how to mine the subtopics of a given query by indirect and direct subtopic mining methods. Then, how to merge and re-rank the retrieved results based on the mined subtopics for search result diversification is described.

3.1 Subtopic mining

In this section, we describe the subtopic mining methods that mine subtopics of a query for supporting the subtopic-based diversified retrieval system. The query "dinosaur", which was selected from the TREC09 topic set, is taken as an example to illustrate the characteristics of each subtopic mining method.

3.1.1 Clustering-based method

Clustering search results is one of the possible ways to mine subtopics of a given query. Two documents are similar if their representations are similar. Documents of similar representations are grouped together to form a document cluster. The common representation of a document cluster is identified as a subtopic of the query. Given a query, the relevant documents are retrieved, features are extracted from each retrieved document, and the weight of a feature is determined by *tf-idf* as follows:

$$w_{i,d} = (0.5 + \frac{0.5 freq_{i,d}}{\max_d freq}) \times log \frac{N}{n_i}, \quad (1)$$

Table 1 Categories mined by using the GIS and the ODP for the query "dinosaur"

Resource	Category
GIS	Science, Entertainment, Games, Local, Lifestyles, Reference
ODP	Kids and Teens, Shopping, Science, Arts, Society, Games, World, Reference, Regional, Computers, Recreation

where $freq_{i,d}$ is the frequency of feature i in document d, $\max_d freq$ is the maximum feature frequency in document d, N is the total number of retrieved documents, and n_i is the number of documents in which feature i appears.

The k-means clustering algorithm (MacQueen 1967) is performed on the retrieved documents, and the documents of similar representations are put together in a document cluster. In this method, the cosine distance determines the similarity between two documents. The number of clusters (e.g., k) has to be determined before the k-means clustering algorithm is applied. The number of clusters is regarded as the number of subtopics within the given query. Intuitively, the number of subtopics depends on the query itself. For example, the query "Obama family tree" may have three subtopics like "TIME magazine photo essay", "Barack Obama's parents and grandparents come from", and "biographical information on Barack Obama". Another query "kcs" may have five subtopics, such as "Kansas City Southern railroad homepage", "job information with the Kansas City Southern railroad", "Kanawha County Schools in West Virginia homepage", "Knox County School system" and "KCS Energy". Therefore, finding an appropriate k is an important issue.

This paper considers three strategies to determine the k. These strategies are based on empirical observation and two external resources, i.e., "Google Insight for Search[1]" (GIS) and "Open Directory Project[2]" (ODP). For the empirical observation, we set k to 5, 10, and 20 because our observation shows the number of subtopics of a query is unlikely to be more than twenty. Furthermore, we employ external resources to determine the k. The GIS provides information of a query based on users' Internet search patterns on Google, such as search volume patterns across specific regions, geographic distribution, and categories for a query. The GIS classifies queries on the Web into 27 categories. We consult the GIS to collect all possible categories of a given query. The number of categories is regarded as the number of its subtopics. The ODP is constructed and maintained by a vast, global community of volunteer editors. It contains more than four million web pages that are organized into more than 500 thousand categories. All websites in the ODP are classified into sixteen major categories. We submit queries to the ODP to collect the major categories of the returned web pages. The number of distinct major categories is regarded as the number of subtopics.

Table 1 lists the collected categories based on the two external resources for the query "dinosaur". As shown in the table, the query "dinosaur" is classified into six major categories by the GIS and eleven major categories by the ODP. This reflects the fact that the ODP provides more specific categories than the GIS.

After document clustering, critical terms in each document cluster are extracted to generate subtopics. The *tf-idf* values of terms in each document cluster are calculated. The

[1] http://www.google.com/insights/search/.

[2] http://www.dmoz.org/.

Table 2 Subtopics mined by the clustering-based method for the query "dinosaur"

No.	Subtopic	No.	Subtopic
1	Dinosaur museum rex national game	6	Dinosaur party pcs shirt barney
2	Dinosaur bird modern fossil animal	7	Dinosaur point compsognathu number
3	Dinosaur word color illustration art	8	Dinosaur toy rex regular animal
4	Dinosaur result attack stegosauru rex	9	Dinosaur museum fossil dino discovery
5	Dinosaur student paper color word	10	Dinosaur walk museum picture live

top five terms of the highest *tf-idf* values in each document cluster are selected as a subtopic of a query.

Table 2 lists the discovered subtopics by the clustering-based method with $k = 10$ for the query "dinosaur". The mined subtopics demonstrate the user's underlying search intent of the query. For example, users may have a search intent focused on dinosaur museums, tyrannosaurus rex, or dinosaur toys. The subtopics mined from the clustering-based method may be duplicated in the surface form. For example, No. 1, No. 9 and No. 10 subtopics refer to the same search intent of "find museum about dinosaur". After deep analysis, the search intents among the three subtopics are somewhat different. The search intent of No. 1 subtopic is the dinosaur game in a museum, No. 9 subtopic is fossils in a museum, and No. 10 is dinosaur pictures in a museum.

Each subtopic has a weight to represent its importance. The size of each document cluster takes into account the relative importance of a subtopic for a query. In addition, the documents in a sub-ranking list of a subtopic are sorted by the descending order of their original ranking scores from the retrieval model.

3.1.2 Topic-category-based method

For the second method, we employ taxonomic information to discover subtopics from the retrieved documents for a given query. We assume that the retrieved documents are considered to have the same subtopic if they are classified into common categories. The categories of taxonomy are treated as their subtopics. In this paper, we use the AlchemyAPI toolkit[3] to obtain the taxonomic information. The AlchemyAPI is a text mining toolkit that utilizes natural language processing technologies and sophisticated statistical algorithms to analyze the input documents before assigning the most likely topic categories.

The retrieved documents are submitted to the toolkit to acquire their categories along with the confidence scores. All distinct categories are aggregated as the subtopics of the query. Documents of the same category then are grouped together to form a category cluster. Note that a document may belong to more than one category cluster if it is classified into different categories. The size of a category cluster takes into account the relative importance of a subtopic. In addition, the documents in a sub-ranking list of a subtopic are sorted by the descending order of their original ranking scores from the retrieval model multiplied by the confidence score of the category. The confidence score represents the relevance between a category and a document. A category with a higher confidence score is more relevant to the document.

Table 3 lists the discovered subtopics for the query "dinosaur". As dinosaurs are paleontological science, most of the documents are classified in the "science technology"

[3] http://www.alchemyapi.com/.

Table 3 Subtopics mined by the topic-category-based method for the query "dinosaur" in the order of their relative importance

Importance	Subtopic
1	Science technology
2	Business
3	Recreation
4	Arts entertainment
5	Culture
6	Politics
7	Computer internet
8	Health
9	Gaming
10	Religion
11	Sports
12	Law crime
13	Unknown

category. Dinosaur information is described in museum-related documents, thus many documents belong to the "arts entertainment" category. The "sports" category is not related to dinosaur at first glance. After examination, we find a golf course and a basketball team named Dinosaur; consequently, they are classified into the "sports" category. Moreover, some documents are categorized into the "unknown" category if they do not belong to any predefined categories of the AlchemyAPI.

3.1.3 Concept-tag-based method

The tags of a document provide a brief summary of the document. Documents labeled with similar concept tags may infer a sort of similar concepts. Folksonomy created by collaborative tagging and social tagging is a good resource to mine subtopics.

In this method, we also employ the AlchemyAPI toolkit to analyze the retrieved documents and generate concept tags along with their confidence scores. The confidence score represents the relevance between a concept tag and a document. The concept tag of a higher confidence score for a document means the concept tag is more relevant to the document. The generated concept tags are regarded as subtopics in this method. Documents then are grouped to form a concept cluster if they contain the same concept tags. Note that a document may belong to more than one concept cluster if it is labeled with more than one concept tag. The size of a concept cluster is considered as the relative importance of the subtopic. In addition, the documents in a sub-ranking list of a subtopic are sorted by the descending order of their original ranking score from the retrieval model multiplied by the confidence score of the subtopic.

Table 4 lists the discovered subtopics for the query "dinosaur". As shown in the table, most of the subtopics are indeed associated with dinosaur. Some interesting subtopics are generated, such as "Jurassic Park", "Michael Crichton" and "Morrison Formation". All of these subtopics are relevant to dinosaurs: Jurassic Park is an American science fiction adventure film, Michael Crichton is the writer of the Jurassic Park novel, and many dinosaur fossils have been found at Morrison Formation in North America. Comparing Tables 3 and 4, we observe that the subtopics discovered from taxonomic information are more general than those from folksonomy.

Table 4 Subtopics mined by the concept-tag-based method for the query "dinosaur" in the order of relative importance

Importance	Subtopic
1	Dinosaur
2	Jurassic Park
3	Tyrannosaurus
4	Museum
5	Michael Crichton
6	Paleontology
7	Natural History Museum
8	Morrison Formation

3.1.4 Ontology-based method

Queries are ambiguous because query terms may refer to more than one sense. Identifying the senses of queries may be a possible method to uncover the subtopics of queries. Wikipedia has become a source of sense annotations for word sense disambiguation. The Wikipedia disambiguation pages provide a service to predict the correct sense of an input query. After query sense disambiguation, a list of references pages is reported for each sense. Each returned sense is regarded as a subtopic of the query. Take Table 5 as an example for the query "dinosaur". The Wikipedia disambiguation page returns five senses: places, film and television, music, comics, and other uses. For the sense "Film and television", four snippets are reported to describe its meaning. One of them is about a Disney computer animated film named Dinosaur, which was a popular film released in 2000.

For each subtopic, the contents of the related web pages are crawled and merged together to form a pseudo document for the subtopic. The number of related web pages is considered as the relative importance of the subtopic. The cosine similarity between a retrieved document and each pseudo document is computed. A document is assigned to the subtopic of the highest similarity score. The documents in a sub-ranking list of a subtopic are sorted by the descending order of their cosine similarity scores.

3.1.5 Query-logs-based method

Search query logs consisting of large collection of search sessions provide an opportunity to discover subtopics embedded in a query. In a search session, users represent their search intents by queries and URL clicks. Similar representations demonstrate similar search intents. Various types of information are used to represent a session. We consult the ODP to obtain more information for each clicked URL. A URL in the ODP is assigned at least one category path. For example, one of the ODP category paths for the URL "http://www.microsoft.com/" is "Computers/Companies/Microsoft_Corporation". In addition, the ODP provides a textual description for each category path and webpage of URL. For each session, we collect the following information: (1) the query terms, (2) the clicked URLs, (3) the ODP category paths of the clicked URLs, (4) the ODP category path descriptions of the clicked URLs, (5) the webpage descriptions for the clicked URLs, (6) the webpage titles of the clicked URLs, and (7) the clicked URLs' contents to form a pseudo document. All pseudo documents are indexed via the Indri search engine. Given a query, a set of sessions in terms of pseudo documents are returned. We also employ the k-means algorithm to cluster the retrieved sessions, where k is determined by the three

Table 5 Subtopics mined by the ontology-based method for the query "dinosaur"

Places
Dinosaur National Monument, on the border of Colorado and Utah in the United States
Dinosaur, Colorado, a town in the United States
Dinosaur, a museum about dinosaurs, in Espéraza, south of France
Film and television
Dinosaur (film), a 2000 Disney computer animated film
Dinosaurs!—A Fun-Filled Trip Back in Time!, a 1987 children's video
Extreme Dinosaurs, an American animated series from 1997 based on a Mattel toy line
Dinosaur, a 1998 short film
Music
Dinosaur (song), by the band King Crimson
Dinosaur Jr., an American indie rock band, originally known as "Dinosaur"
The Dinosaurs, a band
Comics
Dinosaur Comics, a web comic by Ryan North
Other uses
Dempster Dinosaur, a garbage-handling vehicle made by Dempster Brothers

Table 6 Subtopics mined by the query-log-based method for the query "dinosaur"

No.	Subtopic	No.	Subtopic
1	Dinosaur rex USA click	6	Dinosaur student lesson web
2	Dinosaur bird science fossil	7	Dinosaur fossil picture rex
3	Dinosaur lesson fossil student	8	Dinosaur bed toy kid
4	Dinosaur fossil rex art prehistoric	9	Dinosaur color plant print click
5	Dinosaur game available player check	10	Dinosaur party birthday kid toy

alternative strategies, i.e., empirical observation, the GIS, and the ODP, as mentioned in Sect. 3.1.1. Sessions of similar representations are put into a cluster. In this method, the terms with the highest *tf-idf* values in the pseudo documents of the same cluster are considered a subtopic.

Table 6 lists the discovered subtopics for the query "dinosaur" by the query-logs-based method. Compared with the cluster-based method, the "fossil" related subtopics are duplicated in the query-logs-based method and the "museum" related subtopics are not mined. These phenomena show that users may be less interested in the "museum" related subtopics at the time the search query logs were recorded. Thus, the query-logs-based method is a time sensitive approach that captures only the interesting subtopics of users during the search query logs record time.

Recall that each session is represented as a pseudo document. All of the sessions in the same cluster are merged together to form a pseudo document set and used to represent a subtopic of the query. The cosine similarity between a retrieved document and each pseudo document set is computed. A document is classified into the subtopic of the highest similarity score. The documents in a sub-ranking list of a subtopic are sorted by the descending order of their similarity scores. The size of a cluster (e.g., the number of

sessions in a cluster) is taken into account in deciding the relative importance of the subtopic.

3.1.6 Related-search-based method

Most commercial search engines provide the related search mechanisms based on their up-to-date users' search query logs, which record users' long-term searching and query formulation behaviors. The related searching mechanism provides external knowledge sources for subtopic mining. Through such a mechanism, related search queries are expanded from the original query. In this way, an expanded query describes an information need more precisely, based on global user search behaviors recorded in the query logs. Given a query, we collect the related search queries and each related search query is regarded as a subtopic. We utilize three major commercial search engines (e.g., Google, Yahoo, and Bing) for the reference searches. The retrieved documents are assigned to some subtopic(s) as follows. After obtaining the related search query (i.e., subtopics), we query the search engine again using the related search queries. Each subtopic gets a subtopic-ranking list. The number of documents reported by the search engine is taken into account in determining the relative importance of a subtopic. URLs of the original retrieved results for a given query are matched with those URLs in each subtopic-ranking list. If a URL appears in the corresponding subtopic-ranking list, then the document is assigned to the subtopic. Since some URLs may not be covered by any subtopic-ranking list, we propose an approximate approach to deal with the problem. This approach condenses an uncovered URL one level at a time and looks up the subtopic-ranking lists to check if the condensed URL exists in any subtopic-ranking lists. If it exists, the document is assigned to the subtopic. Otherwise, we condense one more level until either a match is found or a miss is reported. Note that a document may be placed into more than one subtopic. The documents in a sub-ranking list of a subtopic are placed by the ranks of the matched URLs.

Table 7 lists the subtopics for the query "dinosaur" mined by the related-search-based methods of the three major commercial search engines. The subtopics mined from Google are shorter and simpler than Yahoo and Bing after observation. In contrast to the previous subtopic mining methods, the number of duplicate subtopics is decreased. The subtopics mined from different search engines are not exactly the same, but the concepts are similar. For example, the subtopic "Play Dinosaur Games" mined from Bing is similar to

Table 7 Subtopics mined by the related-search-based method for the query "dinosaur"

Yahoo	Bing	Google
Walking with dinosaurs	Clip Art Dinosaurs	Dinosaurs types
Pictures of dinosaurs	Dinosaur Coloring Pages	Dinosaurs video
Dinosaurs TV show	Types of Dinosaurs	Dinosaurs names
Dinosaurs for kids	Facts About Dinosaurs	Dinosaurs show
Ice age dawn of the dinosaurs	Play Dinosaur Games	Dinosaurs list
Dinosaurs in the bible	Dinosaurs That Lived in Water	Dinosaurs pictures
Types of dinosaurs	Dinosaurs in the Bible	Dinosaur games
Jurassic park 4 trailer dinosaurs	Dinosaur Specie	Dinosaur facts
Dinosaurs videos		
Imagine dinosaur		

"dinosaur games" mined from Google. Nevertheless, there is no game-related subtopic mined from Yahoo. As a result, the related-search-based method takes into account various search engines to achieve the complementary effects.

3.2 Document ranking

Following the subtopic mining for a given query, two document ranking algorithms are applied to the retrieved results for achieving the goal of search result diversification.

3.2.1 Round-robin for diversification

The round-robin algorithm (abbreviated as RR hereafter) is the simplest merging approach to integrate results of various retrieval models. The main idea is to select the highest relevant documents from each subtopic in a specific order then combine the selected documents as a final ranking list. The subtopics are ordered by their relative importance. Given a query q, there are four major steps in the RR-based diversification algorithm.

- Retrieve the top n documents by Indri search engine.
- Discover the subtopics of q and place the n documents into an appropriate sub-ranking list by the respective subtopic mining methods.
- Arrange the subtopics in the descending order of their relative importance.
- Select the most relevant document from each sub-ranking list in a round way from the most important subtopic to the less important one.

3.2.2 Subtopics for diversification

In this algorithm, the search result diversification is formulated as an optimization problem that aims at optimizing an objective function with regard to the relevance and the diversity. As selecting an optimum document set is an NP-hard problem (Carterette 2009), we use a greedy algorithm that integrates the clues of the subtopics mined from the aforementioned techniques to maximize the objective function. The objective function F is defined as follows:

$$F(q, D) = \rho Rel(q, D) + (1 - \rho) Div(q, D) \tag{2}$$

where D is a set of retrieved documents for a given query q and a parameter $\rho \in [0, 1]$ affects the extent of diversification. If the parameter ρ is equal to 1, then the retrieved results will be ranked in terms of relevance scores only. In contrast, the diversity scores will be only considered if the parameter ρ is set to 0. The objective function is based on a $Rel(q, D)$ function, which estimates the relevance of D with respect to q, and a $Div(q, D)$ function, which measures the diversity of D for q.

Next, a greedy algorithm starts with an empty document set $D' = \varnothing$ and iteratively selects a document d in an attempt to maximize the objective function. The selection process can be defined as follows:

$$d = \arg\max_{d \in D} (F(q, D \cup d) - F(q, D)). \tag{3}$$

The document that maximizes the objective function is added to document set D' until the number of documents in D' meets a predefined threshold.

Below, we describe how the selection process works. The objective function, as shown in Eq. (2), consists of two functions: $Rel(q, D)$ and $Div(q, D)$. We use the $Rel(q, d)$ function

to measure a relevance score of a document $d \in D$, and we use the $Div(q, d)$ function to estimate a diversity score of the document $d \in D$. Previous studies (Carbonell and Goldstein 1998; Yin et al. 2009) compute the relevance score of documents by the existing conventional retrieval models, such as the language model and the vector space model. The limitation of these strategies is that relevance scores of the same document determined by different retrieval models may be unable to be compared. Although relevance scores of documents are usually real numbers, they may be in different ranges, on different scales, and in different distributions. Therefore, the relevance score may dominate the objective function if it is too large, and this may not be effective for diversifying search results. In this paper, the relevance score of a document is computed and normalized as the reciprocal of the rank of a ranking list for generating a comparable relevance score between different retrieval models. Given a document ranking list for q and a document $d \in D$, the $Rel(q, D)$ function is defined as follows:

$$\begin{aligned} Rel(q, D) &= \sum_{d \in D} Rel(q, d) \\ &= \sum_{d \in D} \frac{1}{rank(q, d)} \end{aligned} \tag{4}$$

where $rank(q, d)$ returns the rank of document d in the ranking list for q.

We consider three dimensions for calculating the diversity function $Div(q, D)$. The three components include the *richness* of subtopics, the *importance* of subtopics, and the *novelty* of subtopics within the retrieved documents. Let $Sub(q)$ denote a set of mined subtopics for a given query q and D' denote a document set selected from D. The $Div(q, D)$ function is defined as follows:

$$\begin{aligned} Div(q, D) &= \sum_{d \in D} Div(q, d) \\ &= \sum_{d \in D} richness_d\big(importance_{s \in Sub(q)}(s, q) \cdot novelty_{s \in Sub(q), D' \subset D}(s, D')\big) \end{aligned} \tag{5}$$

where $richness_d(\cdot)$ measures the richness of subtopics covered by the document d, $importance(s, q)$ measures the relative importance of the subtopic s, and $novelty(s, D')$ measures the novelty of the subtopic s in the selected document set D'. In other words, the number of subtopics that d covers, the number of important subtopics that d contains and the number of novelty subtopics that d includes is directly related to the diversity score that d has.

As mentioned before, the subtopics are mined by various subtopic mining methods, and a document belonging to more than one subtopic is more likely to satisfy more users. Thus, the richness function ranks the documents of broad subtopics, i.e., documents with more subtopics move to the top positions of a ranking list. Furthermore, the richness function considers clues from subtopics mined by various subtopic mining methods. The richness function merges the coverage of subtopics mined by different subtopic mining methods by summing the covered subtopics and averaging by the number of subtopic mining methods. We rewrite the $Div(q, d)$ function in Eq. (5) as follows:

$$Div(q, d) = \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \big(importance_{s \in S(q)}(s, q) \cdot novelty_{s \in Sub(q), D' \subset D}(s, D')\big) \tag{6}$$

where m is the number of subtopic mining methods used and n_i is the number of subtopics mined by the subtopic mining method i.

As for the relative importance of a subtopic, we incorporate the factor into the $Div(q, d)$ function directly. The relative importance of subtopic s depends on different subtopic mining methods, as mentioned in Sect. 3.1. With the incorporation, Eq. (6) can be rewritten as follows:

$$Div(q,d) = \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(weight_{i,j} \cdot novelty_{s\in Sub(q),D'\subset D}(s,D')\right) \tag{7}$$

where $weight_{i,j}$ is a relative importance of the j-th subtopic generated by the subtopic mining method i.

Since some evaluation metrics, like α-nDCG, are designed based on the novelty-biased cumulative gain, the α-nDCG score is higher if the documents at top ranks of a ranking list contain different subtopics. Inspired by α-nDCG, we design the *novelty* function to highlight top-position documents containing different subtopics. The *novelty* function gives a penalty to a document if it has the same set of subtopics as those documents that have already been selected in D'. With the concept of α-nDCG, the $Div(q, d)$, shown in Eq. (7), can be rewritten as follows:

$$Div(q,d) = \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(weight_{i,j}\frac{1}{rank(d,s_{i,j})}(1-\alpha)^{\sum_{d'\in D'}\frac{1}{rank(d',s_{i,j})}}\right) \tag{8}$$

where $rank(d, s_{i,j})$ is the rank of document d in the $s_{i\ j}$ sub-ranking list, $s_{i,j}$ is the j-th subtopic generated by subtopic mining method i, D' is the set of documents which have been selected, and α is a parameter for penalizing duplicate subtopics, $\alpha \in [0, 1]$. In the following experiments, α is set to 0.5. Different from the previous studies, which mainly focus on one aspect of diversification, we deal with the aspects of *richness*, *importance*, and *novelty* together in an object function for diversifying search results.

4 Experimental resources

In this paper, the ClueWeb09 dataset[4] is the main experimental resource. The ClueWeb09 dataset is a standard Web collection developed by the text retrieval conference TREC09 Web Track (Clarke et al. 2009), where the dataset is divided into two collections, i.e., the Category A and the Category B. The Category A collection contains the entire dataset of 5 billion English pages, and the Category B collection selects the first 50 million English pages. We use the Indri search engine,[5] which is a popular academic information retrieval toolkit, to index the corpus. Before indexing, the documents are stemmed using the Porter stemmer and the stop words are filtered.

The MSN Search Query Log excerpt (RFP 2006 dataset) (Craswell et al. 2009b) is an important resource for subtopic mining. It consists of 14.9 million queries and 12.2 million clicks during a one-month period in May 2006. The MSN Search Query Log excerpt is separated into two files: one named query and the other named click. The query file is described by a set of attributes, including Time, Query, QueryID, and ResultCount, and the click file contains attributes, such as QueryID, Query, Time, URL, and URL Position. Note that these two files are linked through QueryID. In total, there are 7.4 million sessions, which contain the activities of a user from the time of the first query submission to the time of a timeout between the web browser and the search engine. The contents of the clicked URLs are crawled in the search query logs.

[4] http://lemurproject.org/clueweb09/.

[5] http://www.lemurproject.org/indri/.

```
<topic number="1" type="faceted">
  <query>obama family tree</query>
  <description>
    Find information on President Barack Obama's family
    history, including genealogy, national origins,
    places and dates of birth, etc.
  </description>
  <subtopic number="1" type="nav">
    Find the TIME magazine photo essay "Barack Obama's
    Family Tree".
  </subtopic>
  <subtopic number="2" type="inf">
    Where did Barack Obama's parents and grandparents
    come from?
  </subtopic>
  <subtopic number="3" type="inf">
    Find biographical information on Barack Obama's
    mother.
  </subtopic>
</topic>
```

Fig. 2 An example topic along with its corresponding subtopics in the TREC09

National Institute of Standards and Technology (NIST) created and assessed 50 topics for each diversity task in the TREC09 and TREC10 Web tracks. As shown in Fig. 2, each topic contains a query field, a description field, and several subtopic fields, but only the query field was released to participants. For each topic, participants in the diversity tasks submitted a ranking of the top 10,000 documents for that topic. All submitted runs were included in the pool for judgment. In this paper, our experiments are conducted on the ClueWeb09 dataset Category B and Category A test collections using the topics of the TREC09 and TREC10 Web Track diversity tasks.

5 Experimental results and discussion

In this section, we first evaluate the performance of the different subtopic mining methods we propose. Then, we compare the two proposed diversification algorithms, i.e., the RR-based diversification algorithm and the subtopic-based diversification algorithm. We use topics on the TREC09 and TREC10 Web Track for testing, and we experiment on the ClueWeb09 Category B and Category A test collections. In the experiments, three metrics are used for evaluation. Finally, we compare our best model with the state-of-the-art models in the TREC09 and TREC10 Web Track diversity tasks and discuss the experimental results.

Experiments were evaluated by three well-known metrics: (1) α-nDCG (Clarke et al. 2008), which measures the overall relevance across intents; (2) Intent-Aware Precision (Agrawal et al. 2009) (abbreviated as IA-P), which measures the diversity; and (3) α#-nDCG, which is a linear combination of IA-P and α-nDCG. Assume there are k subtopics for a given query. The α-nDCG metric considers the k subtopics for computing the gain vector. The α-nDCG is defined as:

$$\alpha - nDCG[k] = \frac{\alpha - DCG[k]}{\alpha - DCG'[k]} \tag{9}$$

where α-$DCG[k]$ is a normalized discounted cumulative gain vector and α-$DCG'[k]$ is the ideal discounted cumulative gain vector.

Intent-aware precision (IA-P) at retrieval depth l is defined as follows:

$$IA - P@l = \frac{1}{M}\sum_{t=1}^{M}\frac{1}{N_t}\sum_{i=1}^{N_t}\frac{1}{l}\sum_{j=1}^{l} j_t(i,j) \tag{10}$$

where M is the number of topics. Let N_t ($1 \leq t \leq M$) be the number of subtopics associated with topic number t. Let $j_t(i,j) = 1$ if the document returned for topic t at depth j is judged relevant to subtopic i of topic t; otherwise 0. We list the performance at depths of $l = 5$, 10, and 20 (i.e., the number of top ranked items to be evaluated), in order to comprehensively examine the effect of the proposed systems.

5.1 Evaluation of the subtopic mining methods

A total of 100 topics, along with several example subtopics, were provided in the TREC09 and TREC10 Web Track diversity tasks. The average number of example subtopics per topic is 4.83 and 4.36 for the TREC09 and TREC10, respectively. The example subtopics for each topic were regarded as a ground truth for this topic. We conducted a user study to evaluate the performance of the proposed subtopic mining techniques. The assessors were asked to identify which mined subtopics were relevant to the TREC ground truth. Additionally, they also identified the subtopics that did not appear in the ground truth, but were still relevant to the test topics. To avoid bias, the mined subtopics were presented to the assessors randomly, so they did not know which subtopics were generated by which subtopic mining method.

The subtopic mining performance on the TREC09 is shown in Table 8. The clustering-based method performed the best among the indirect subtopic mining methods, where k was set to 10 for the k-means clustering algorithm. The performance of the best method is highlighted in bold in Table 8 and the subsequent tables. This approach achieved α#-nDCG@3 0.406, α#-nDCG@5 0.377, and α#-nDCG@10 0.324. The performance using the GIS and the ODP to determine the number of clusters was lower than that of using the empirical strategy. The average number of subtopics determined by the GIS and the ODP were 4.16 and 8.86, respectively. When the number of clusters is decreased, documents with different subtopics may be placed in the same cluster, so some subtopics may not be generated. The topic-category-based and the concept-tag-based subtopic mining methods may also suffer from the same problem.

The related-search-based methods used the up-to-date users' search query logs from three commercial search engines (i.e., Bing, Yahoo, and Google). The method using Bing performed the best among the direct subtopic mining methods. The ontology-based method used the Wikipedia disambiguation page for mining subtopics. Nevertheless, only 17 topics had the corresponding disambiguation pages in the 50 test topics. The overall performance was decreased by the remaining 33 test topics. Similarly, the performance of the query-logs-based model depended on the search query log dataset. The logging time affected the mining results. Take the topic "Obama family tree" as an example. The MSN Search Query Log excerpt reflects users searching and browsing information in 2006. Since Obama was not the president of the United States at that time, there is little information in the search query logs. We hardly can mine subtopics from time-dependent queries using out-of-date search query logs. The ground truth is another issue. Take the topic "dinosaurs" as an example. The example subtopics in the ground truth contain (1) Discovery Channel dinosaur pictures and games, (2) free pictures of dinosaurs, (3) pictures of dinosaurs that can be colored in, (4) different kinds of dinosaurs pictures, and (5) homepage BBC series Walking Dinosaurs. After observation, some additional subtopics mined

Table 8 Performance of the subtopic mining methods on the TREC09 using the ClueWeb09 Category B test collection

TREC09	α-nDCG			IA-P			α#-nDCG		
	@3	@5	@10	@3	@5	@10	@3	@5	@10
Indirect mining methods									
Topic-Category	0.391	0.333	0.290	0.095	0.075	0.052	0.243	0.204	0.171
Concept-tag	0.490	0.443	0.439	0.146	0.144	0.130	0.318	0.294	0.285
Clustering (GIS)	0.535	0.437	0.365	0.142	0.112	0.062	0.339	0.275	0.214
Clustering (ODP)	0.565	0.524	0.514	0.175	0.174	0.154	0.370	0.349	0.334
Clustering ($k = 10$)	**0.625**	**0.577**	**0.520**	**0.186**	**0.177**	**0.127**	**0.406**	**0.377**	**0.324**
Direct mining methods									
Ontology	0.342	0.292	0.258	0.088	0.072	0.046	0.215	0.182	0.152
Query logs (GIS)	0.370	0.328	0.273	0.089	0.070	0.038	0.230	0.199	0.156
Query logs ($k = 10$)	0.416	0.405	0.391	0.114	0.118	0.099	0.265	0.262	0.245
Query logs (ODP)	0.426	0.390	0.354	0.105	0.096	0.070	0.266	0.243	0.212
Related search (Google)	0.589	0.574	0.545	0.163	0.169	0.132	0.376	0.372	0.339
Related search (Yahoo)	0.648	0.601	0.571	0.183	0.174	0.134	0.416	0.388	0.353
Related search (Bing)	**0.662**	**0.612**	**0.603**	**0.183**	**0.181**	**0.160**	**0.423**	**0.397**	**0.382**

from the query-logs-based method are related to the topic, but do not appear in the ground truth. The descendants or ancestors of dinosaurs and dinosaur movies are typical examples.

Table 9 lists the performance of subtopic mining methods on the TREC10. The tendency is similar to the performance shown in Table 8. To sum up, the direct subtopic mining methods perform more effectively than the indirect subtopic mining methods.

Since the subtopics of a TREC topic are just examples for the topic, some additional subtopics mined by the subtopic mining methods may be related to the test topic, but do not appear in the ground truth. To clarify this point, we compare the strict and the lenient performance of the subtopic mining methods. The strict performance is based on the original ground truth by the TREC only, and the lenient performance considers those additional subtopics that are regarded as correct by the assessors. Figures 3 and 4 show the strict and the lenient performance of different subtopic mining methods on the TREC09 and TREC10, respectively. Intuitively, the lenient performance is better than the strict performance. Take the ontology-based method as an example. The ontology-based method achieves a significant improvement because more than half of the additional subtopics are discovered by the ontology-based method, compared with the number of subtopics in the TREC09 and TREC10 ground truth after observation.

5.2 Baseline models for document ranking

In the document ranking experiments, four models were conducted as baselines, including the initial Indri retrieval model[6] (*IRM*), the maximal marginal relevance (*MMR*) model (Carbonell and Goldstein 1998), the language modeling approach *WUME* (Yin et al. 2009), and the explicit query aspect diversification (*xQuAD*) model (Santos et al. 2010a). As mentioned in Sect. 2.2, *MMR* is categorized as the implicit diversification approach and *WUME* and *xQuAD* are classified into the explicit diversification approach. The *WUME*

[6] http://ciir.cs.umass.edu/~metzler/indriretmodel.html.

Table 9 Performance of the subtopic mining methods on the TREC10 using the ClueWeb09 Category B test collection

TREC10	α-nDCG			IA-P			α#-nDCG		
	@3	@5	@10	@3	@5	@10	@3	@5	@10
Indirect mining methods									
Topic-Category	0.288	0.292	0.297	0.088	0.095	0.093	0.188	0.194	0.195
Concept-tag	0.307	0.290	0.277	0.086	0.079	0.067	0.197	0.185	0.172
Clustering (GIS)	0.491	0.473	0.445	0.154	0.153	0.123	0.323	0.313	0.284
Clustering (ODP)	0.507	0.441	0.369	0.160	0.129	0.068	0.334	0.285	0.219
Clustering ($k = 10$)	**0.503**	**0.488**	**0.484**	**0.164**	**0.162**	**0.156**	**0.334**	**0.325**	**0.320**
Direct mining methods									
Ontology	0.292	0.265	0.254	0.069	0.061	0.045	0.181	0.163	0.150
Query logs ($k = 10$)	0.451	0.451	0.448	0.136	0.144	0.140	0.294	0.298	0.294
Query logs (GIS)	0.494	0.442	0.378	0.137	0.113	0.064	0.316	0.278	0.221
Query logs (ODP)	0.505	0.482	0.456	0.159	0.156	0.118	0.332	0.319	0.287
Related search (Yahoo)	0.626	0.600	0.560	0.193	0.188	0.138	0.410	0.394	0.349
Related search (Google)	0.710	0.684	0.657	0.213	0.202	0.154	0.462	0.443	0.406
Related search (Bing)	**0.719**	**0.664**	**0.644**	**0.221**	**0.207**	**0.184**	**0.470**	**0.436**	**0.414**

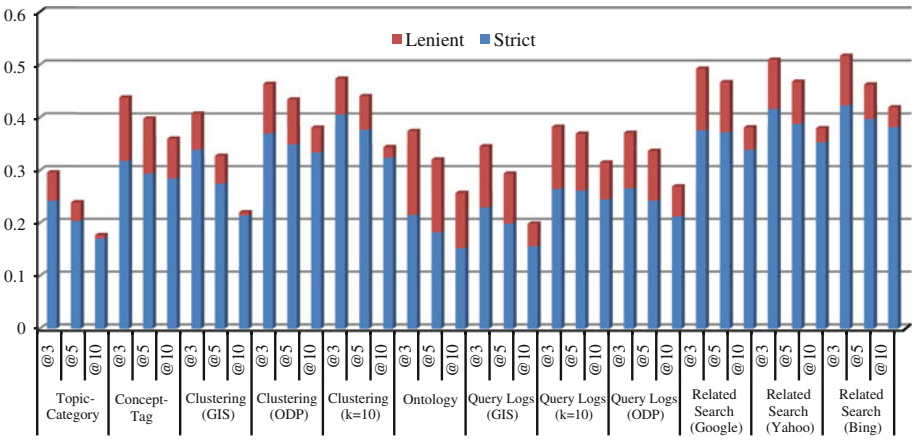

Fig. 3 The strict and lenient performance of the subtopic mining methods on the TREC09

and *xQuAD* approaches use the subtopics explicitly mined by the related-search-based method using Bing. We also formulated the baseline models as an optimization problem, and the greedy algorithm was used to optimize both the relevance function and the diversity function. The relevance function was the same as ours, as shown in Eq. (4). The diversity functions of the four baseline models are described as follows.

(1) *IRM* model is the Indri initial retrieval model that combines the language modeling (Ponte and Croft 1998) and inference network (Turtle and Croft 1991) with Dirichlet smoothing (Zhai and Lafferty 2004). That is, there is no diversity function in the *IRM* model.

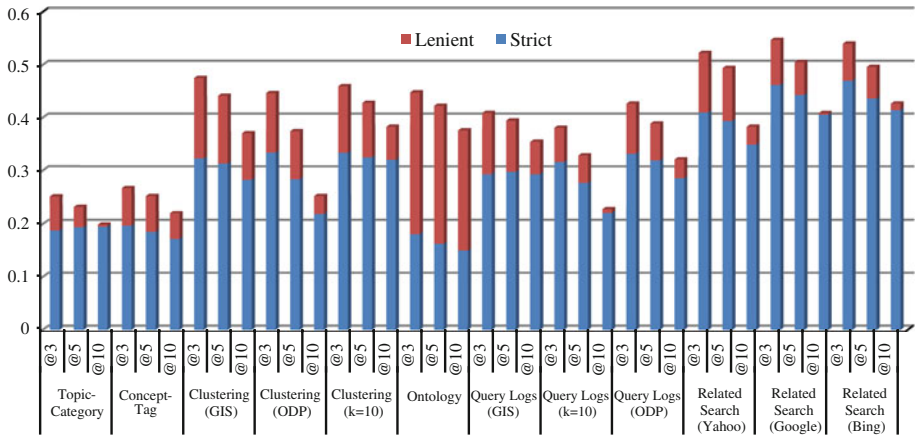

Fig. 4 The strict and lenient performance of the subtopic mining methods on the TREC10

(2) *MMR* model optimizes the margin relevance of the documents and iteratively selects the document that has high marginal relevance if it is relevant to the query and contains minimal similarity to the previously selected documents. The diversity function of the model is defined as follows.

$$Div_{MMR}(q,d,D) = -\max_{d' \in D'} p(d|d') \tag{11}$$

(3) *WUME* model maximizes the probability in such a way that the document meets the user search intents in terms of subtopics through selecting documents that maximally cover the subtopics of a query. The diversity function of the model is defined as follows.

$$Div_{WUME}(q,d,D) = \sum_{s \in S(q)} p(s|q)p(d|s) \tag{12}$$

(4) *xQuAD* model uses a probability model to diversify search results as well. The model not only optimizes the coverage of subtopics, but also gives a penalty to a document covering the subtopics that have been well covered by the previously selected documents. The diversity function of the model is defined as follows:

$$Div_{xQuAD}(q,d,D) = \sum_{s \in S(q)} p(s|q)p(d|s) - \prod_{d' \in D'} (1 - p(d'|s)) \tag{13}$$

where D is a set of retrieved documents for a given query q, D' is the set of previously selected documents, $S(q)$ is the subtopic set of q, $p(d|d')$ measures the likelihood of a document $d \in D$ and the selected document $d' \in D'$, $p(d|s)$ measures the likelihood of the document d and the subtopic $s \in S(q)$, and $p(s|q)$ measures the likelihood of the subtopic s and the query q.

5.3 Parameter determination

To determine a parameter k of the empirical strategy in the clustering-based and the query-logs-based subtopic mining methods, along with a parameter ρ in the subtopic-based diversification algorithm, we performed 5-fold cross-validation over the fifty topics on both the TREC09 and TREC10, optimizing the evaluation metric α#-nDCG@5. The parameter

k was regarded as the number of clusters to represent the number of subtopics. The parameter ρ affects how diversification is considered. If the parameter ρ is equal to 1, then the ranking score will be generated according to the relevance part only. Conversely, the diversity part is considered only if the parameter ρ is set to 0. In other words, we put more weight on the diversity part to rank the initial retrieved documents if ρ is smaller than 0.5. The interpolation parameter ρ employed by the three baselines, i.e., *MMR*, *WUME*, and *xQuAD*, was determined in the same way to balance both the relevance and the diversity.

5.4 Evaluation on the ClueWeb09 Category B collection

Tables 10 and 11 list the performance of the four baseline models and the RR-based diversification algorithm with subtopics mined from various subtopic mining methods on the TREC09 and TREC10, respectively, using the ClueWeb09 Category B test collection. The tendency of the two tables is similar. The *IRM* model is the weakest baseline. The experimental results show that *MMR*, *WUME*, and *xQuAD* perform better than *IRM* in both Tables 10 and 11. This is because the Indri initial retrieval model concentrates on retrieving relevant documents and underestimates the search result diversification. The baselines with the explicit approach, i.e., *WUME* and *xQuAD*, making use of subtopics mined by the related-search-based method with Bing, are more effective than *MMR*. The *xQuAD* model performs the best among the four baseline models. The α#-nDCG@5, α#-nDCG@10, and α#-nDCG@20 of *xQuAD* on the TREC09 are 0.141, 0.172, and 0.195, respectively. In addition, the α#-nDCG@5, α#-nDCG@10, and α#-nDCG@20 of *xQuAD* on the TREC10 are 0.137, 0.161, and 0.180, respectively. In the Wilcoxon signed rank tests, *xQuAD* performs significantly better than *IRM* and *MMR* on α#-nDCG at all measurement depths ($p < 0.05$). This reflects that using the explicit approach is better than the implicit approach, which is consistent with the study (Santos et al. 2010a). Nevertheless, the performance of *xQuAD* cannot significantly outperform *WUME*.

For the performance of the RR-based diversification algorithm, as shown in Table 10, the clustering-based method ($k = 10$) performs the best among the indirect subtopic mining methods. It achieves the best α#-nDCG@5 0.186, significantly outperforming the strongest baseline (i.e., *xQuAD*) about by 27 percent and outperforming the other three baselines ($p < 0.05$). Using the GIS and the ODP to determine the number of clusters is lower than the empirical strategy. The average number of subtopics determined by the GIS and the ODP is 4.16 and 8.86, respectively. If the number of clusters is decreased, then documents of different subtopics may be placed in the same cluster. In such a case, the RR-based diversification algorithm will miss some subtopics if documents belong to those subtopics at lower ranks. The topic-category-based and the concept-tag-based subtopic mining methods suffer from the same problem.

For direct subtopic mining methods, the related-search-based models using the up-to-date users' search query logs from three commercial search engines (e.g., Bing, Yahoo, and Google) to generate subtopics are the most effective to diversify the search results. The experiments show that their performance is very similar. The related-search-based method using Bing, which performs the best among the direct subtopic mining methods, is significantly better than *IRM* and *MMR* (p value < 0.05). The tendency of the RR-based diversification algorithm on the TREC10 Category B test collection is the same as that on the TREC09 Category B test collection, except that the best indirect and direct subtopic mining methods with the RR-based diversification algorithm significantly outperform all four baselines ($p < 0.05$).

Table 10 Performance of the RR-based diversification algorithm on the TREC09 using the ClueWeb09 Category B test collection

TREC09	α-nDCG			IA-P			α#-nDCG		
	@5	@10	@20	@5	@10	@20	@5	@10	@20
Baseline									
IRM	0.148	0.195	0.237	0.079	0.091	0.096	0.114	0.140	0.167
MMR	0.151	0.199	0.240	0.081	0.094	0.097	0.116	0.147	0.169
WUME	0.183	0.231	0.274	0.082	0.096	0.100	0.133	0.164	0.187
xQuAD	**0.196**	**0.246**	**0.287**	**0.085**	**0.097**	**0.103**	**0.141**	**0.172**	**0.195**
Indirect mining method +RR-based diversification									
Topic-Category	0.180	0.222	0.262	0.09	0.098	0.100	0.135	0.160	0.181
Clustering (GIS)	0.205	0.242	0.278	0.099	0.098	0.096	0.152	0.170	0.187
Concept-tag	0.214	0.240	0.281	0.094	0.085	0.085	0.154	0.163	0.183
Clustering (ODP)	0.226	0.251	0.287	0.096	0.093	0.092	0.161	0.172	0.190
Clustering ($k = 10$)	**0.256**	**0.272**	**0.307**	**0.116**	**0.100**	**0.097**	**0.186**	**0.186**	**0.202**
Direct mining method +RR-based diversification									
Query logs (GIS)	0.165	0.204	0.248	0.079	0.086	0.093	0.122	0.145	0.171
Query logs (ODP)	0.178	0.212	0.255	0.082	0.084	0.089	0.130	0.148	0.172
Ontology	0.191	0.212	0.252	0.083	0.079	0.081	0.137	0.146	0.167
Query logs ($k = 10$)	0.191	0.222	0.269	0.083	0.082	0.093	0.137	0.152	0.181
Related search (Google)	0.199	0.227	0.266	0.094	0.096	0.100	0.147	0.162	0.183
Related search (Yahoo)	0.202	0.242	0.270	0.092	0.102	0.101	0.147	0.172	0.186
Related search (Bing)	**0.204**	**0.237**	**0.271**	**0.098**	**0.097**	**0.099**	**0.151**	**0.167**	**0.185**

To sum up, the performance of the RR-based diversification algorithm is highly dependent on the subtopic mining performance. As shown in Table 8 and Table 9, the clustering-based method ($k = 10$) and the related-search-method using Bing achieve the best performance of the indirect and direct subtopic mining methods, respectively. This meets our expectation because the RR-based diversification algorithm is based on the mined subtopics to re-rank the retrieved documents. The more accurately subtopics are identified; the more diversified results are generated.

As mentioned above, six different subtopic-mining methods are proposed. The subtopic-based diversification algorithm integrates clues of various subtopics mined by these methods. Table 12 lists the experimental results on the TREC09 using the ClueWeb09 Category B test collection, where DC($k = 10$) denotes the cluster-based method and k is set to a fixed number 10 in the k-means clustering algorithm, CT denotes the concept-tag-based method, TC denotes the topic-category-based method, RS(B) denotes the related-search-based method using Bing, RS(Y) denotes the related-search-based method using Yahoo, RS(G) denotes the related-search-based method using Google, and RS(ALL) denotes the integration of the related search results of Bing, Yahoo, and Google.

The performance of the top ten combinations in terms of α#-nDCG@5 is shown in Table 12. The related-search-based methods using Google combined with Bing or Yahoo achieve the best performance of α#-nDCG@5 0.200, which is significantly better than the strongest baseline, i.e., *xQuAD* ($p < 0.05$). The up-to-date users' search query logs from commercial search engines are very important data sources for the subtopic-based diversification algorithm.

Table 11 Performance of the RR-based diversification algorithm on the TREC10 using the ClueWeb09 Category B test collection

TREC10	α-nDCG			IA-P			α#-nDCG		
	@5	@10	@20	@5	@10	@20	@5	@10	@20
Baseline									
IRM	0.135	0.172	0.212	0.089	0.096	0.090	0.112	0.134	0.151
MMR	0.139	0.176	0.217	0.090	0.097	0.093	0.115	0.137	0.155
WUME	0.167	0.205	0.242	0.093	0.101	0.095	0.130	0.153	0.169
xQuAD	**0.179**	**0.218**	**0.263**	**0.095**	**0.103**	**0.096**	**0.137**	**0.161**	**0.180**
Indirect mining method +RR-based diversification									
Topic-Category	0.178	0.220	0.259	0.092	0.090	0.089	0.135	0.155	0.174
Concept-tag	0.193	0.220	0.264	0.100	0.092	0.090	0.147	0.156	0.177
Clustering (GIS)	0.190	0.222	0.256	0.113	0.108	0.102	0.152	0.165	0.179
Clustering (ODP)	0.205	0.223	0.271	0.101	0.079	0.090	0.153	0.151	0.181
Clustering ($k = 10$)	**0.250**	**0.272**	**0.298**	**0.128**	**0.105**	**0.077**	**0.189**	**0.189**	**0.188**
Direct mining method +RR-based diversification									
Query logs (GIS)	0.139	0.180	0.228	0.080	0.087	0.097	0.110	0.134	0.163
Ontology	0.172	0.202	0.223	0.081	0.080	0.066	0.127	0.141	0.145
Query logs ($k = 10$)	0.174	0.207	0.254	0.087	0.086	0.100	0.131	0.147	0.177
Query logs (ODP)	0.187	0.213	0.254	0.100	0.093	0.097	0.144	0.153	0.176
Related search (Yahoo)	0.227	0.265	0.305	0.128	0.115	0.106	0.178	0.190	0.206
Related search (Google)	0.256	0.283	0.319	0.141	0.114	0.101	0.199	0.199	0.210
Related search (Bing)	**0.250**	**0.285**	**0.317**	**0.148**	**0.131**	**0.114**	**0.199**	**0.208**	**0.216**

Integration of subtopics mined by other approaches, such as the topic-category-based and the concept-tag-based, does not improve the performance significantly.

The subtopic-based diversification model is also better than the best RR-based diversification model, i.e., the RR-based diversification algorithm with subtopics mined by the Clustering ($k = 10$) subtopic mining method. The RR-based algorithm does not consider a document to be classified into more than one subtopic, so the performance may be decreased. In contrast, the subtopic-based diversification algorithm not only prefers documents with multiple subtopics (e.g., *richness*), but also considers the *importance* and *novelty* of subtopics, as mentioned at Sect. 3.2.2. Table 13 lists the experimental results on the TREC10 using the ClueWeb09 Category B test collection, whose tendency is similar to those on the TREC09 counterpart.

The proposed diversification algorithm is based on an initial retrieved document set, so the initial results will affect the performance of search result diversification. Table 14 shows the experimental results of the subtopic-based diversification algorithm using initial document sets retrieved by different academic open-source search engines, including Indri,[7] a search engine toolkit designed by University of Massachusetts and Carnegie Mellon University, and Zettair,[8] developed by the search engine group at Royal Melbourne Institute of Technology University (RMIT). Both search engines are reported to have good performance (Middleton and Baeza-Yates 2007) on information retrieval. We employed

[7] Indri: http://www.lemurproject.org/indri/.

[8] Zettair: http://www.seg.rmit.edu.au/zettair/.

Table 12 Performance of the subtopic-based diversification algorithm on the TREC09 using the ClueWeb09 Category B test collection

TREC09	α-nDCG			IA-P			α#-nDCG		
	@5	@10	@20	@5	@10	@20	@5	@10	@20
The best baseline (*xQuAD*)	0.196	0.246	0.287	0.085	0.097	0.103	0.141	0.172	0.195
The best RR-based model	0.256	0.272	0.307	0.116	0.100	0.097	0.186	0.186	0.202
RS(ALL) + DC($k = 10$) + CT + TC	0.269	0.286	0.326	0.108	0.084	0.083	0.189	0.185	0.205
RS(ALL) + DC($k = 10$) + TC	0.269	0.283	0.328	0.110	0.082	0.086	0.190	0.183	0.207
RS(ALL) + DC($k = 10$)	0.269	0.283	0.329	0.111	0.081	0.087	0.190	0.182	0.208
RS(ALL) + CT + TC	0.276	0.304	0.335	0.109	0.096	0.085	0.193	0.200	0.210
RS(B) + RS(Y)	0.277	0.311	0.344	0.109	0.101	0.098	0.193	0.206	0.221
RS(ALL)	0.281	0.317	0.348	0.111	0.102	0.098	0.196	0.210	0.223
RS(ALL) + TC	0.283	0.316	0.347	0.108	0.101	0.090	0.196	0.209	0.219
RS(ALL) + CT	0.284	0.307	0.343	0.111	0.097	0.086	0.198	0.202	0.215
RS(B) + RS(G)	0.285	0.305	0.337	0.114	0.100	0.095	0.200	0.203	0.216
RS(G) + RS(Y)	**0.280**	**0.301**	**0.338**	**0.119**	**0.100**	**0.095**	**0.200**	**0.201**	**0.217**

three retrieval models on the Indri search engine, including the language model (LM), the language model with pseudo-relevance feedback (LM + PRF), and Okapi BM25. For the Zettair search engine, the language model (LM) and Okapi BM25 were employed.

As shown in Tables 12 and 13, the best model achieved 0.200 and 0.269 in terms of α#-nDCG@5 on the TREC09 and TREC10 using the ClueWeb09 Category B test collection, respectively, under the condition that the initial document set was retrieved by the Indri search engine with the language model. When the initial document set was retrieved by the Indri search engine with the language model and pseudo-relevance feedback, i.e., Indri (LM + PRF); the performance of α#-nDCG@5 improved to 0.214 (7 % improvement) and 0.304 (13 % improvement) on the TREC09 and TREC10, respectively. For the Indri search engine, using the language model with pseudo-relevance feedback is better than using the language model only and the Okapi BM25 model. With the Zettair search engine, using the language model is better than using the Okapi BM25 model. The tendency of the retrieval models is similar on the two academic open-source search engines. To sum up, if the initial document set covers more relevant documents, the models will be able to diversify search results with high effectiveness.

We also compare the best proposed diversification model with the three state-of-the-art models in the TREC09 Web Track diversity task (Clarke et al. 2009), as shown in Table 15. The best model we proposed uses Indri with the language model and pseudo-relevance feedback to retrieve an initial document set and diversifies the initial document set via the subtopic-based diversification algorithm with subtopics mined by the related-search-based method using Google and Yahoo. The Amsterdam team used the Latent Dirichlet Allocation (LDA) to extract 10 topics from the top 2,500 documents in the initial retrieved document set and represented each document as a mixture of 10 topics. Inspired by MMR, their diversity function maximized the expected joint probability of all topics in the selected result set for search result diversification (He et al. 2009). The ICTNET team used the *k*-means clustering algorithms to cluster the retrieved documents and diversify these documents based on the size of clusters (Bi et al. 2009). The uogTr team used query

Table 13 Performance of the subtopic-based diversification algorithm on the TREC10 using the ClueWeb09 Category B test collection

TREC10	α-nDCG			IA-P			α#-nDCG		
	@5	@10	@20	@5	@10	@20	@5	@10	@20
The best baseline (*xQuAD*)	0.179	0.218	0.263	0.095	0.103	0.096	0.137	0.161	0.180
The best RR-based model	0.250	0.285	0.317	0.148	0.131	0.114	0.199	0.208	0.216
RS(ALL) + DC($k = 10$)	0.294	0.341	0.377	0.130	0.12	0.103	0.212	0.231	0.240
RS(ALL) + DC($k = 10$) + TC	0.292	0.331	0.371	0.140	0.125	0.116	0.216	0.228	0.244
RS(ALL) + DC($k = 10$) + CT + TC	0.293	0.311	0.362	0.139	0.108	0.097	0.216	0.210	0.230
RS(ALL) + CT + TC	0.294	0.333	0.365	0.143	0.122	0.102	0.219	0.228	0.234
RS(ALL)	0.299	0.337	0.370	0.147	0.123	0.100	0.223	0.230	0.235
RS(ALL) + CT	0.308	0.322	0.371	0.141	0.102	0.095	0.225	0.212	0.233
RS(ALL) + TC	0.302	0.330	0.369	0.150	0.119	0.101	0.226	0.225	0.235
RS(B) + RS(Y)	0.312	0.348	0.387	0.147	0.128	0.117	0.230	0.238	0.252
RS(B) + RS(G)	0.338	0.367	0.402	0.167	0.131	0.113	0.253	0.249	0.258
RS(G) + RS(Y)	**0.358**	**0.379**	**0.408**	**0.179**	**0.135**	**0.110**	**0.269**	**0.257**	**0.259**

Table 14 Comparison of different academic open-source search engines

Retrieval search engine	α-nDCG			IA-P			α#-nDCG		
	@5	@10	@20	@5	@10	@20	@5	@10	@20
TREC09									
Zettair (okapi BM25)	0.260	0.287	0.316	0.108	0.097	0.080	0.184	0.192	0.198
Indri (Okapi BM25)	0.272	0.303	0.332	0.120	0.103	0.091	0.196	0.203	0.212
Indri (LM)	0.280	0.301	0.338	0.119	0.100	0.095	0.200	0.201	0.217
Zettair (LM)	0.288	0.311	0.336	0.113	0.100	0.085	0.201	0.206	0.211
Indri (LM + PRF)	**0.307**	**0.340**	**0.380**	**0.121**	**0.106**	**0.097**	**0.214**	**0.223**	**0.239**
TREC10									
Zettair (okapi BM25)	0.324	0.349	0.384	0.167	0.131	0.109	0.246	0.240	0.247
Indri (Okapi BM25)	0.333	0.358	0.397	0.167	0.131	0.110	0.250	0.245	0.254
Indri (LM)	0.358	0.379	0.408	0.179	0.135	0.110	0.269	0.257	0.259
Zettair (LM)	0.373	0.394	0.422	0.179	0.138	0.113	0.276	0.266	0.268
Indri (LM + PRF)	**0.395**	**0.421**	**0.457**	**0.212**	**0.201**	**0.167**	**0.304**	**0.311**	**0.312**

suggestions from a major Web search engine as subtopics of a given topic and applied *xQuAD* for search result diversification. The symbols *, +, and ▲ in Table 15 show that the improvement over the three state-of-the-art models, the Amsterdam, the ICTNET, and the uogTr, respectively, are statistically significant ($p < 0.05$). As shown in Table 15, our proposed model significantly outperforms the Amsterdam and the ICTNET on α#-nDCG@5 and α#-nDCG@10. In addition, our proposed model is significantly better than the three state-of-the-art models on α#-nDCG@20.

We next discuss the possible reasons for the improvement of our model over the state-of-the-art models. The *MMR* baseline is similar to the model of the Amsterdam team. The main difference between the *MMR* baseline and their approach is that the subtopics are

Table 15 Comparison of the state-of-the-art models proposed in the TREC09 on the ClueWeb09 Category B test collection

TREC09	α-nDCG			IA-P			α#-nDCG		
	@5	@10	@20	@5	@10	@20	@5	@10	@20
Amsterdam	0.232	0.250	0.281	0.086	0.079	0.071	0.159	0.165	0.176
ICTNET	0.251	0.272	0.301	0.104	0.095	0.092	0.178	0.184	0.197
uogTr	0.253	0.282	0.308	0.142	0.132	0.127	0.198	0.207	0.218
Our model	0.307	0.340	0.380	0.121	0.106	0.097	$0.214^{*,+}$	$0.223^{*,+}$	$0.239^{*,+,\blacktriangle}$

implicitly extracted by LDA and are encoded in their model. Our model is better than the Amsterdam team because our diversity function combines subtopics mined by two direct methods. The subtopics mined by the direct mining method are generally better than the indirect method, which is also demonstrated from the experimental results shown in Tables 8 and 9. The RR-based diversification algorithm with subtopics mined from the clustering-based method is similar to the model of the ICTNET team. Our model performs better than the ICTNET team because their model does not consider the case where a document is classified into more than one subtopic and because the subtopics they used also are mined by the implicit method. The *xQuAD* baseline model is the same as the diversification algorithm of the uogTr team. Nevertheless, the initial document set and the subtopics used are different. The reasons our model performs better than the uogTr team may be due to the initial document retrieval performance being improved. Moreover, our diversity function combines subtopics mined by different aspects and considers the relative importance of subtopics.

Table 16 lists the experimental results of the best diversification model we proposed and the three state-of-the-art models in the TREC10 Web Track diversity task (Clarke et al. 2010) for comparison. In the TREC10, the performance of α-nDCG@5 and IA-P@5 were not reported. Moreover, the performance was computed over 36 of the 50 topics. The held back topic numbers were 54, 61, 66, 68, 72, 78, 83, 86, 87, 90, 95, 98, 99, and 100. The performance of the best model we proposed is better than the three state-of-the-art models in the TREC10. For the UAmsterdam team, the documents were retrieved based on the similarity scores between the anchor texts of documents and a given query and were ranked by the similarity scores multiplied by the fusion spam percentiles (Kamps et al. 2010). The uogTr team used the same model as that in the TREC09, but the subtopics were generated based on query reformulations of Bing and Google (Santos et al. 2010c). Unfortunately, the framework of the qirdcsuog team was not reported in the TREC10 proceedings. Since the performance of each test topic for the three top models was not reported in the TREC10, the results shown in Table 16 are listed without significance tests.

We further discuss the possible reasons our model is better than the three state-of-the-art models. Our model is better than the UAmsterdam team because they only used the information from anchor texts to retrieve the initial document set. The initial retrieval performance affects the performance of search result diversification significantly, as we discussed above. The uogTr team used a similar approach as they employed in the TREC09. The possible reasons are reported above.

5.5 Evaluation on the ClueWeb09 Category A collection

Tables 17 and 18 list the experimental results of the subtopic-based diversification algorithm on the ClueWeb09 Category A collection using the TREC09 and TREC10 topics,

Table 16 Comparison of the state-of-the-art models proposed in the TREC10 on the ClueWeb09 Category B test collection

TREC10	α-nDCG		IA-P		α#-nDCG	
	@10	@20	@10	@20	@10	@20
UAmsterdam	0.272	0.319	0.111	0.102	0.191	0.210
qirdcsuog	0.269	0.302	0.154	0.140	0.212	0.221
uogTr	0.399	0.451	0.184	0.169	0.292	0.310
Our model	0.421	0.457	0.201	0.167	0.311	0.312

respectively. The performance of the four baseline models on the ClueWeb09 Category A test collection is lower than that on Category B test collection, as shown in Tables 10 and 11, because the number of documents of the ClueWeb09 Category A test collection is ten times larger than the Category B test collection. Nevertheless, the performance tendency of the four baseline models on the both TREC09 and TREC10 is similar, i.e., $IRM < MMR < WUME < xQuAD$. The *xQuAD* model performs significantly better than *IRM* and *MMR* on α#-nDCG at all measurement depths ($p < 0.05$). For the performance of the top ten combinations in terms of α#-nDCG@5 shown in Table 17, the related-search-based method using the combination of Google, Bing, and Yahoo achieves the best performance of α#-nDCG@5 0.164, α#-nDCG@10 0.164, and α#-nDCG@20 0.166, which is significantly better than all of the baselines at all measurement depths ($p < 0.05$). This reflects again that the up-to-date users' search query logs from commercial search engines are very

Table 17 Performance of the subtopic-based diversification algorithm on the TREC09 using the ClueWeb09 Category A test collection

TREC09	α-nDCG			IA-P			α#-nDCG		
	@5	@10	@20	@5	@10	@20	@5	@10	@20
Baseline									
IRM	0.045	0.071	0.090	0.020	0.031	0.034	0.033	0.051	0.062
MMR	0.053	0.079	0.097	0.021	0.033	0.036	0.037	0.056	0.067
WUME	0.079	0.105	0.123	0.025	0.037	0.039	0.052	0.071	0.081
xQuAD	**0.095**	**0.122**	**0.139**	**0.026**	**0.038**	**0.041**	**0.061**	**0.080**	**0.090**
Subtopic-based diversification									
RS(ALL) + QL	0.211	0.228	0.258	0.075	0.059	0.052	0.143	0.144	0.155
RS(ALL) + QL + DC($k = 10$) + CT	0.202	0.215	0.237	0.086	0.065	0.051	0.144	0.140	0.144
RS(ALL) + DC($k = 10$) + TC	0.202	0.212	0.235	0.087	0.066	0.053	0.145	0.139	0.144
RS(ALL) + DC($k = 10$) + CT	0.202	0.213	0.237	0.087	0.065	0.054	0.145	0.139	0.146
RS(ALL) + DC($k = 10$) + QL	0.205	0.216	0.239	0.085	0.063	0.051	0.145	0.140	0.145
RS(ALL) + CT + TC	0.208	0.229	0.257	0.084	0.067	0.055	0.146	0.148	0.156
RS(ALL) + DC($k = 10$)	0.214	0.226	0.250	0.088	0.068	0.057	0.151	0.147	0.154
RS(ALL) + CT	0.218	0.240	0.270	0.088	0.071	0.059	0.153	0.156	0.165
RS(ALL) + TC	0.226	0.243	0.266	0.094	0.071	0.058	0.160	0.157	0.162
RS(ALL)	**0.233**	**0.251**	**0.271**	**0.095**	**0.077**	**0.061**	**0.164**	**0.164**	**0.166**

Table 18 Performance of the subtopic-based diversification algorithm on the TREC10 using the ClueWeb09 Category A test collection

TREC10	α-nDCG			IA-P			α#-nDCG		
	@5	@10	@20	@5	@10	@20	@5	@10	@20
Baseline									
IRM	0.039	0.061	0.086	0.028	0.034	0.043	0.034	0.048	0.065
MMR	0.044	0.067	0.093	0.029	0.036	0.046	0.037	0.052	0.070
WUME	0.064	0.087	0.113	0.029	0.036	0.046	0.047	0.062	0.080
xQuAD	**0.085**	**0.108**	**0.135**	**0.034**	**0.039**	**0.050**	**0.060**	**0.074**	**0.093**
Subtopic-based diversification									
RS(ALL) + QL	0.246	0.269	0.290	0.119	0.101	0.078	0.183	0.185	0.184
RS(ALL) + QL + DC($k = 10$) + CT	0.256	0.282	0.307	0.127	0.101	0.095	0.192	0.192	0.201
RS(ALL) + DC($k = 10$) + QL	0.269	0.292	0.324	0.121	0.095	0.081	0.195	0.194	0.203
RS(ALL) + DC($k = 10$) + CT	0.268	0.291	0.312	0.132	0.105	0.085	0.200	0.198	0.199
RS(ALL) + DC($k = 10$) + TC	0.290	0.315	0.343	0.124	0.102	0.087	0.207	0.209	0.215
RS(ALL) + DC($k = 10$)	0.280	0.306	0.332	0.135	0.114	0.097	0.208	0.210	0.215
RS(ALL) + CT + TC	0.295	0.319	0.349	0.127	0.104	0.093	0.211	0.212	0.221
RS(ALL) + CT	0.295	0.319	0.349	0.127	0.104	0.093	0.211	0.212	0.221
RS(ALL) + TC	0.292	0.311	0.332	0.144	0.110	0.093	0.218	0.211	0.213
RS(ALL)	**0.326**	**0.343**	**0.365**	**0.158**	**0.123**	**0.100**	**0.242**	**0.233**	**0.233**

important data sources to mine subtopics for the subtopic-based diversification algorithm. The tendency of the experimental results on the TREC10 (i.e., Table 18) is similar to that on the TREC09 counterpart (i.e., Table 17). The performance of the best model is significantly better than all of the baselines at all measurement depths ($p < 0.05$) as well.

As mentioned before, the performance of the initial retrieval document set is a key factor for the search result diversification. We employ the Indri search engine with the language model and pseudo-relevance feedback to retrieve an initial document set, and we diversify the initial document set via the subtopic-based diversification algorithm with subtopics mined by the related-search-based method using Bing, Google, and Yahoo. We further compare the best proposed diversification model with the three state-of-the-art models in the TREC09 Web Track diversity task (Clarke et al. 2009), as shown in Table 19. The THUIR team used the BM25 retrieval model to retrieve relevant documents and clustered these documents. Each cluster was taken as a probable subtopic, and the IA-SELECT algorithm (Agrawal et al. 2009) was employed for diversifying search results (Li et al. 2009). The msrc team proposed a retrieval system that considered three features, BM25 score, PageRank, and the matching anchor count, and estimated ranking scores for documents by linear combination of the three factors. The weights of the three features were trained by their own search engine logs. Then, they diversified the top ranks documents based on "host collapsing" (Craswell et al. 2009a). The MSRAsia team proposed a search result diversification algorithm that used subtopics mined from anchor texts, search results clusters, and sites of search results. They used the BM25 retrieval model to retrieve an initial document set and employed a greedy algorithm to iteratively select the best document from this set to maximize the coverage of subtopics for search result

Table 19 Comparison of the state-of-the-art models proposed in the TREC09 on the ClueWeb09 Category A test collection

TREC09	α-nDCG			IA-P			α#-nDCG		
	@5	@10	@20	@5	@10	@20	@5	@10	@20
THUIR	0.206	0.234	0.271	0.105	0.094	0.086	0.156	0.164	0.179
msrc	0.268	0.309	0.346	0.127	0.117	0.105	0.198	0.213	0.226
MSRAsia	0.281	0.316	0.365	0.127	0.112	0.108	0.204	0.214	0.237
Our Model	0.314	0.350	0.367	0.127	0.111	0.083	$0.221^{*,+}$	0.231*	0.225*

Table 20 Comparison of the state-of-the-art models proposed in the TREC10 on the ClueWeb09 Category A test collection

TREC10	α-nDCG		IA-P		α#-nDCG	
	@10	@20	@10	@20	@10	@20
THUIR	0.426	0.475	0.156	0.138	0.291	0.306
msrsv	0.428	0.478	0.186	0.171	0.307	0.325
uwgym	0.453	0.500	0.181	0.177	0.317	0.339
Our Model	0.427	0.445	0.209	0.163	0.318	0.304

diversification (Dou et al. 2009). The symbols *, +, and ▲, in Table 19 show that the improvement over the three state-of-the-art models, the THUIR, the msrc and the MSRAsia, respectively, is statistically significant ($p < 0.05$). As shown in Table 15, the best model we proposed significantly outperforms than the THUIR on α#-nDCG at all measurement depths and the msrc on α#-nDCG@5.

We next discuss the possible reasons for the improvement of our model over the state-of-the-art models. The used subtopics of the THUIR team were indirectly extracted by document clustering. Our model is better than the THUIR team because our diversity algorithm combines subtopics mined by the direct mining methods. The model of the msrc team concentrated on relevant documents retrieved, and diversified search results only considered the host names of the retrieved documents. Our model performs better than the msrc team because our proposed diversification algorithm not only keeps the quality of relevancy, but also re-ranks the retrieved documents to cover multiple and important subtopics. The MSRAsia used the BM25 retrieval model to retrieve the initial document sets. Our model performs better than the MSRAsia team because of the initial document set and the parameter optimization. As shown in Table 14, the initial document set by the language model and pseudo-relevance feedback is better than that using the BM25 retrieval model. In addition, the parameters of the object function in the subtopic-based diversification algorithm are optimized.

Table 20 lists the experimental results of the best model we proposed and the three state-of-the-art models in the TREC10 Web Track diversity task (Clarke et al. 2010) for comparison. The performance of our proposed model is better than the three state-of-the-art models on α#-nDCG@10 in the TREC10. The possible reasons our model is better than the three state-of-the-art models are similar to the discussions above.

6 Conclusion and future work

In this paper, we propose six subtopic mining methods for mining subtopics from different aspects and present two document ranking algorithms to diversify search results with the mined subtopics. To reduce the number of missing subtopics and keep fewer redundant subtopics, subtopics are mined indirectly from the retrieved documents or directly from queries themselves. The proposed subtopic-based diversification algorithm considers the *richness*, the *importance*, and the *novelty* of the mined subtopics together Experimental results show the subtopic-based diversification algorithm can balance both the relevance and the diversity for improving the performance of search result diversification.

We analyzed the effectiveness of the subtopics derived from different subtopic mining methods. A user study of comparing the subtopics mined by different mining methods with the ground truth in the TREC09 and TREC10 is explored. A thorough evaluation of our proposed diversification models within the standard experiment paradigm in the TREC09 and TREC10 Web Tracks diversity tasks is conducted. A set of experiments is carried out to verify the effectiveness of the proposed diversification models. Experimental results show that the best model uses the Indri search engine with the language model and pseudo-relevance feedback to retrieve an initial document set, and diversifies the initial document set via subtopics mined by the related-search-based method with the subtopic-based diversification algorithm. The best model outperforms most of the state-of-the-art models proposed in the TREC09 and TREC10 Web Track diversity tasks significantly.

We also found that the initial retrieval performance of the retrieval models is important for the diversification of the search results. The initial document set covering documents that are more relevant can provide more information for mining subtopics and increase the effectiveness of search result diversification. The subtopics granularity of a query affects the performance of search result diversification as well. For the cluster-based methods, using the empirical strategy to determine the number of subtopics is better than consulting external resources, such as the GIS and the ODP. The query-logs-based model is highly time-dependent for subtopic mining. The performance of the query-logs-based model may be out of expectation if the related information is not recorded in the duration of logs. The related-search-based method applying the up-to-date users' search query logs to generate subtopics confirms the results. Compared with other subtopic mining methods, subtopics generated by the related-search-based method have good quality and less duplication. The performance of the subtopic-based diversification algorithm is better than the RR-based diversification algorithm.

Future directions include how to integrate other knowledge resources into the diversification models further, such as social information, and how to extend this work to diversify Web search results with different languages. How to localize the diversified models to meet users' needs from different areas/countries also has to be dealt with in the future. Moreover, we plan to detect duplicate subtopics after the subtopic mining phrase. Detection of duplicate subtopics is expected to refine the subtopic mining performance and then enhance the performance of search result diversification.

Acknowledgments This work was partially supported by National Science Council (Taiwan) and Excellent Research Projects of National Taiwan University under contracts NSC98-2221-E-002-175-MY3, NSC99-2221-E-002-167-MY3, and 101R890858.

References

Agrawal, R., Gollapudi, S., Halverson, A., & Ieong, S. (2009). Diversifying search results. In *Proceedings of the 2nd ACM International Conference on Web Search and Data Mining* (pp. 5–14).

Bi, W., Yu, X., Liu, Y., Guan, F., Peng, Z., Xu, H., & Cheng, X. (2009). ICTNET at Web Track 2009 diversity track. In *Proceedings of the 18th Text REtrieval Conference*.

Boldi, P., Bonchi, F., Castillo, C., Donato, D., Gionis, A., & Vigna, S. (2008). The query-flow graph: Model and applications. In *Proceeding of the 17th ACM Conference on Information and Knowledge Management* (pp. 609–618).

Broder, A. (2002). A taxonomy of web search. *SIGIR Forum, 36*(2), 3–10.

Carbonell, J., & Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 335–336).

Carterette, B. (2009). An analysis of NP-completeness in novelty and diversity ranking. In *Proceedings of the 2nd International Conference on Theory of Information Retrieval: Advances in Information Retrieval Theory* (pp. 200–211).

Carterette, B., & Chandar, P. (2009). Probabilistic models of ranking novel documents for faceted topic retrieval. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management* (pp. 1287–1296).

Chang, Y. S., He, K. Y., Yu, S., & Lu, W. H. (2006). Identifying user goals from Web search results. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence* (pp. 1038–1041).

Chen, H., & Karger, D. R. (2006). Less is more: probabilistic models for retrieving fewer relevant documents. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 429–436).

Clarke, C., Craswell, N., & Soboroff, I. (2009). Overview of the TREC 2009 web track. In *Proceedings of the 18th Text REtrieval Conference.* (pp. 1–9).

Clarke, C. L. A., Craswell, N., Soboroff, I., & Cormack, G. V. (2010). Overview of the TREC 2010 web track. In *Proceedings of the 19th Text REtrieval Conference* (pp. 1–9).

Clarke, C. L. A., Kolla, M., Cormack, G. V., Vechtomova, O., Ashkan, A., Büttcher, S., & MacKinnon, I. (2008). Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 659–666).

Craswell, N., Fetterly, D., Najork, M., Robertson, S., & Yilmaz, E. (2009a). Microsoft Research at TREC 2009: Web and Relevance Feedback Track. In *Proceedings of the 18th Text REtrieval Conference*.

Craswell, N., Jones, R., Dupret, G., & Viegas, E. (2009b). In *Proceedings of the 2009 Workshop on Web Search Click Data* (pp. 95).

Dou, Z., Chen, K., Song, R., Ma, Y., Shi, S., & Wen, J. R. (2009). Microsoft research Asia at the web track of TREC 2009. In *Proceedings of the 18th Text REtrieval Conference*.

Geng, X., Liu, T. Y., Qin, T., Arnold, A., Li, H., & Shum, H. Y. (2008). Query dependent ranking using K-nearest neighbor. In *Proceedings of the 31st a Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 115–122).

Gollapudi, S., & Sharma, A. (2009). An axiomatic approach for result diversification. In *Proceedings of the 18th International Conference on World Wide Web* (pp. 381–390).

He, J., Balog, K., Hofmann, K., Meij, E., Rijke, M. de, Tsagkias, M., & Weerkamp, W. (2009). Heuristic ranking and diversification of web documents. In *Proceedings of the 18th Text REtrieval Conference*.

Hu, J., Wang, G., Lochovsky, F., Sun, J., & Chen, Z. (2009). Understanding user's query intent with wikipedia. In *Proceedings of the 18th International Conference on World Wide Web* (pp. 471–480).

Kamps, J., Kaptein, R., & Koolen, M. (2010). Using anchor text, spam filtering and wikipedia for web search and entity ranking. In *Proceedings of the 19th Text REtrieval Conference*.

Li, Z., Chen, F., Xing, Q., Miao, J., Xue, Y., Zhu, T., Zhou, B., (2009). Thuir at trec 2009 web track: Finding relevant and diverse results for large scale web search. In *Proceedings of the 18th Text REtrieval Conference*.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Math, Statistics, and Probability* (Vol. 1, pp. 281–297).

Manshadi, M., & Li, X. (2009). Semantic tagging of web search queries. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP* (pp. 861–869).

Middleton, C., & Baeza-Yates, R. (2007). Technical report: A comparison of open source search engines. Retrieved from http://wrg.upf.edu/WRG/dctos/Middleton-Baeza.pdf.

Nguyen, V., & Kan, M. Y. (2007). Functional faceted web query analysis. In *Query Log Analysis: Social and Technological Challenges. A workshop at the 16th International World Wide Web Conference*.

Ponte, J. M., & Croft, W. B. (1998). A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 275–281).

Radlinski, F., Bennett, P. N., Carterette, B., & Joachims, T. (2009). Redundancy, diversity and interdependent document relevance. *SIGIR Forum, 43*(2), 46–52.

Radlinski, F., & Dumais, S. (2006). Improving personalized web search using result diversification. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 691–692).

Radlinski, F., & Joachims, T. (2005). Query chains: Learning to rank from implicit feedback. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining* (pp. 239–248).

Rafiei, D., Bharat, K., & Shukla, A. (2010). Diversifying web search results. In *Proceedings of the 19th International Conference on World Wide Web* (pp. 781–790).

Rose, D. E., & Levinson, D. (2004). Understanding user goals in web search. In *Proceedings of the 13th International Conference on World Wide Web* (pp. 13–19).

Santos, R. L. T., Macdonald, C., & Ounis, I. (2010a). Exploiting query reformulations for web search result diversification. In *Proceedings of the 19th International Conference on World Wide Web* (pp. 881–890).

Santos, R. L. T., Macdonald, C., & Ounis, I. (2010b). Selectively diversifying web search results. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management* (pp. 1179–1188).

Santos, R. L. T., McCreadie, R. M. C., Macdonald, C., & Ounis, I. (2010c). University of Glasgow at TREC 2010: Experiments with terrier in blog and web tracks. In *Proceedings of the 19th Text REtrieval Conference*.

Song, R., Zhang, M., Sakai, T., Kato, M. P., Liu, Y., Sugimoto, M., Wang, Q. (2011). Overview of the NTCIR-9 INTENT Task. In *Proceedings of the 9th NTCIR Workshop Meeting*.

Spärck-Jones, K., Robertson, S. E., & Sanderson, M. (2007). Ambiguous requests: implications for retrieval tests, systems and theories. *SIGIR Forum, 41*(2), 8–17.

Turtle, H., & Croft, W. B. (1991). Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems (TOIS), 9*(3), 187–222.

Vee, E., Srivastava, U., Shanmugasundaram, J., Bhat, P., & Yahia, S. A. (2008). Efficient computation of diverse query results. In *Proceedings of the 24th IEEE International Conference on Data Engineering* (pp. 228–236).

Wang, J., & Zhu, J. (2009). Portfolio theory of information retrieval. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 115–122).

Welch, M. J., Cho, J., & Olston, C. (2011). Search result diversity for informational queries. In *Proceedings of the 20th International Conference on World Wide Web* (pp. 237–246).

Yin, D., Xue, Z., Qi, X., & Davison, B. D. (2009). Diversifying search results with popular subtopics. In *Proceedings of the 18th Text REtrieval Conference*.

Yue, Y., & Joachims, T. (2008). Predicting diverse subsets using structural SVMs. In *Proceedings of the 25th International Conference on Machine Learning* (pp. 1224–1231).

Zhai, C. X., Cohen, W. W., & Lafferty, J. (2003). Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval* (pp. 10–17).

Zhai, C., & Lafferty, J. (2004). A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems (TOIS), 22*(2), 179–214.