Hu S, Dou ZC, Wang XJ *et al.* Search result diversification based on query facets. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 30(4): 888–901 July 2015. DOI 10.1007/s11390-015-1567-5

Search Result Diversification Based on Query Facets

Sha Hu (胡 莎), Zhi-Cheng Dou* (窦志成), *Member, CCF, ACM, IEEE*, Xiao-Jie Wang (王晓捷), and Ji-Rong Wen (文继荣), *Senior Member, CCF, ACM, IEEE*

School of Information, Renmin University of China, Beijing 100872, China
Key Laboratory of Data Engineering and Knowledge Engineering, Ministry of Education, Beijing 100872, China

E-mail: {shahu, dou, wangxiaojie}@ruc.edu.cn; jirong.wen@gmail.com

Received February 1, 2015; revised April 24, 2015.

Abstract In search engines, different users may search for different information by issuing the same query. To satisfy more users with limited search results, search result diversification re-ranks the results to cover as many user intents as possible. Most existing intent-aware diversification algorithms recognize user intents as subtopics, each of which is usually a word, a phrase, or a piece of description. In this paper, we leverage query facets to understand user intents in diversification, where each facet contains a group of words or phrases that explain an underlying intent of a query. We generate subtopics based on query facets and propose faceted diversification approaches. Experimental results on the public TREC 2009 dataset show that our faceted approaches outperform state-of-the-art diversification models.

Keywords query intent, query facet, search result diversification

1 Introduction

In search engines, users like to get the most relevant results that match what they want by inputting the easiest queries. Even issuing the same query, different users may want different information to be retrieved. This situation appears more often in short queries, which are usually ambiguous or broad to specify a user's intent[1-4]. Table 1 shows an example of manually edited user intents (or subtopics) for the query (or topic) "Defender" (topic number 20) in TREC 2009[5]. The ambiguous query "defender" has multiple interpretations, including the software "Windows Defender", the car "Land Rover Defender", the marine outfitter "Defender Marine", the game "Defender Arcade Game", and the newspaper "Chicago Defender Newspaper". Even for the clearly defined interpretation "Windows Defender", users may still seek various aspects, such as looking for "windows defender homepage", or finding "windows defender review". Therefore, to satisfy more users, the optimal search results should contain the documents with respect to different intents of the query at the top of the result list.

Table 1. Subtopics of Query "Defender" in TREC 2009 Web Track

No.	Type	Subtopic Description
1	nav	I'm looking for the *homepage* of *Windows Defender*, an anti-spyware program.
2	inf	Find information on the *Land Rover Defender* sport-utility vehicle.
3	nav	I want to go to the homepage for *Defender Marine* supplies.
4	inf	I'm looking for information on Defender, an *arcade game* by Williams. Is it possible to play it online?
5	inf	I'd like to find *user reports* about *Windows Defender*, particularly problems with the software.
6	nav	Take me to the homepage for the *Chicago Defender newspaper*.

Regular Paper

Special Section on Data Management and Data Mining

This work was partially supported by the National Basic Research 973 Program of China under Grant No. 2014CB340403, the Fundamental Research Funds for the Central Universities of China, and the Research Funds of Renmin University of China under Grant No. 15XNLF03.

*Corresponding Author

To solve the above problem, there are two challenges: recognizing query intents and diversifying search results based on query intents. Search result diversification is more focused on promoting diversity for search results, utilizing predefined query intents mined by intent mining approaches. Some public tasks have been organized for search result diversification in the information retrieval community, such as the TREC Web Track Diversity task① and the NTCIR IMine task②. The goal is to return a ranked list of documents that completely cover query intents and avoid excessive redundancy.

In diversification, query intents are implicitly or explicitly used to promote diversity in various ways. For example, MMR[6] measures query intents by document similarity in content. IA-Select[7] makes use of a taxonomy to classify query categories. xQuAD[8] leverages query reformulations from search engines as subtopics. ACSL[9] combines subtopics mined by different approaches from different data sources. DSPApprox[10] organizes subtopics by extracting key terms from search results. For most existing diversification algorithms, each query intent is typically represented as a word, a phrase, or a piece of description, which is a traditional form in intent mining approaches.

In this paper, we move a small step towards utilizing better query intents in diversification. More specifically, we generate faceted subtopics based on query facets[11-13]. The original query facets are designed to help improve the search user experience such as faceted search and exploratory search. Each facet contains a group of words or phrases extracted from search results, which explains an underlying intent of the query. Table 2 shows an example of top five refined subtopics based on query facets for query "Olympic". In traditional diversification, these facets may be represented by subtopics such as "Olympic Sports", "Olympic Countries", "Olympic Colors", "different Olympic Games", and "Olympic Host Cities" respectively. In diversification, using traditional subtopics may misunderstand documents: if a document does not mention the word "Sports", it will be viewed as irrelevant to "Olympic Sports", even if it introduces all the sports in Olympics. Furthermore, if two documents are both related to "Olympic Sports", traditional diversification algorithms may not know which one is better, even if one document only mentions some kind of sports while the other document details all the sports.

Table 2. Faceted Subtopics of Query "Olympic"

No.	Faceted Subtopics
1	Athletics, Boxing, Wrestling, Basketball, Shooting, Triathlon, Football, Swimming, Weightlifting, Archery, Cycling, Gymnastics, Judo, Fencing, Rowing, Volleyball, Equestrian, Handball, Hockey, Tennis, Modern Pentathlon, Badminton, Canoe, Baseball, Taekwondo
2	Greece, France, Britain, United States, Switzerland, Belgium, Italy, Australia, Japan, Germany, Sweden, Mexico, Canada, Japan Norway, Norway, the Soviet Union, Spain, Austria, in the Netherlands, Yugoslavia
3	Blue, yellow, black, green, dark blue, red, sky blue
4	Summer Olympic Games, Winter Olympic Games, Paralympic Games
5	Athens, Paris, Tokyo, Saint Louis, Mexico City, London, Munich, Berlin, Stockholm, Moscow, Amsterdam, Antwerp

Therefore, we exploit query facets in diversification and propose faceted diversification approaches. In particular, we extend state-of-the-art diversification algorithms, viz., MMR[6], IASelect[7], and ACSL[9], and propose a faceted MMR model (FMMR), a faceted IASelect model (FIASelect), and a faceted ACSL model (FACSL). According to the faceted items, we know the real content of each facet in detail. We can precisely estimate the relevance between the documents and the facets, and the relevance between the query and the facets. When selecting the most diverse document, our faceted algorithms tend to select the document that covers the most faceted items of facets that are not covered by the previously selected documents.

We evaluate our approaches on the public document collection ClueWeb09③ and the query set used on TREC 2009 Web track[5]. Experimental results show that by using query facets, the adaptive algorithms can obtain improved diversity compared with their original models, in terms of α-NDCG[14] and Precision-IA[7]. FACSL performs the best in terms of all metrics and it significantly outperforms some state-of-the-art diversification models. The results indicate that exploiting query facets can benefit search result diversification.

The main contributions of this paper are:

• We study the problem of leveraging query facets to generate subtopics.

①TREC. http://plg.uwaterloo.ca/∼trecweb/2012.html, May 2015.

②NTCIR. http://www.thuir.org/imine/, May 2015.

③ClueWeb09. http://boston.lti.cs.cmu.edu/Data/clueweb09/, May 2015.

• We adapt state-of-the-art diversification algorithms, and propose three corresponding faceted models to diversify search results based on faceted subtopics.

• We conduct experiments to demonstrate that faceted subtopics can help improve result diversity.

The remainder of this paper is organized as follows. We briefly discuss related work in Section 2, and introduce how to generate subtopics by query facets in Section 3. Then we propose our faceted diversification algorithms in Section 4, and analyze the experimental results in Section 5. We conclude our work in Section 6.

2 Related Work

The problem of search result diversification has been well studied over a decade. In 1998, Carboness and Goldstein[6] proposed the influential Maximal Marginal Relevance (MMR) algorithm, which attempts to reduce redundancy while maintaining query relevance to high-ranked documents. The novelty and the relevance of search result documents were measured by the similarities of document content, as there was no categorization of either documents or queries at that time. Zhai and Lafferty[15] presented a risk minimization framework for subtopic retrieval, in which relevance and novelty can be modeled together within a loss function[15]. Chen and Karger[16] implemented a blind negative feedback on their probabilistic models to maximize the probability of retrieving one relevant document for a given query. Zhang *et al.*[17] proposed an affinity ranking algorithm to re-rank search results by optimizing diversity and information richness of search results. By assuming that similar documents will cover similar subtopics, the aforementioned approaches implicitly consider the subtopics underlying a query .

Later, Agrawal *et al.*[7] explicitly classified queries and documents based on ODP taxonomy. They proposed a greedy algorithm to maximize the probability of finding at least one useful document in the top results. Santos *et al.*[8] diversified search results based on query reformulations from Web search engines. They also proposed a selective diversification approach to learn a trade-off between relevance and diversity[18], and another learning model to select appropriate retrieval models for different query aspects[19]. Dou *et al.*[9] represented a framework to combine multiple subtopics mined from different data sources. Yue and Joachims[20] learned to predict diverse subsets and maximize result diversity by structural SVMs. Radlinski *et al.*[21] learned to diversify documents by users' click behavior. Rafiei *et al.*[4] treated user clicks as relevance votes, and related result quality and diversity to expected payoff and risk in clicks. Dang and Croft[22] leveraged political election strategy into diversification, and diversified search results by maintaining the proportionality for query aspects. They also used terms as subtopics and proposed term level diversification algorithms[10]. He *et al.*[23] introduced a flexible algorithm to combine multiple external resources. Zhu *et al.*[24] provided a learning-to-rank approach to promote diversity. Yu and Ren[25] treated the diversity task as a multiple subtopic knapsack problem and re-ranked the documents like filling up multiple subtopic knapsacks. Liang *et al.*[26] inferred topic model to get latent subtopics. Although current intent-aware approaches generate query intents from various sources, combinations or models, they commonly represent query intents in a traditional way, where each intent is a word or a phrase. Our work, from another aspect, utilizes a new form of query intents, i.e., facets, each of which is a group of words or phrases that show the real content of the facet.

There have been some approaches to mining topics from documents. For example, Lawrie *et al.*[27] proposed a graph-theoretic algorithm to generate topical hierarchies automatically. Dou *et al.*[11] mined query facets, the multiple groups of words or phrases, to explain the underlying query facets. Hu *et al.*[28] provided a clustering algorithm to mine subtopics from search log data. Kong and Allan[12] developed a supervised approach based on a graphical model to extract query facets from search results. Abbassi *et al.*[29] modeled the diversity maximization problem under matroid constraints. Bache *et al.*[30] used a text-based framework to quantify how diverse a document is in terms of its content. Jameel and Lam[31] discovered topics based on text documents. Although these approaches are designed for different purposes, they all assist information discovery for the query or the documents and can be reasonably used in search result diversification. In this paper, we preliminary refine query facets extracted from search results, and leverage these faceted subtopics to promote diversity. We try to investigate whether our automatically generalized subtopics can better predict user intents and improve result diversity. The detailed analysis of these subtopic mining algorithms is beyond the scope of this work.

3 Generation of Faceted Subtopics

In this paper, we generate faceted subtopics based on query facets, proposed by Dou *et al.*[11] and Kong and Allan[12]. Query facets aggregate frequent lists in search results of a query to explain the underlying query facets. Ideally, query facets can be automatically mined for any query in any open domain. A weight is assigned to each facet, and this might be useful in the later calculations. Recently in the NTCIR-11 IMine Task, this algorithm is adopted to provide subtopic candidates in subtopic mining subtask.

We implement query facets following the framework in [11]. We extract lists from free text, HTML tags, repeat regions from the search results of a query, and group them into clusters based on the items they contain. The output includes multiple facets that summarize the information about the query from multiple perspectives. Each facet is organized in a group of items, including words or phrases. In the framework, facets and their items are evaluated and ranked based on their importance.

However, these original facets cannot be directly adopted as subtopics. Since query facets are designed for splitting different facets of a query, they are usually far more fine-grained than traditional subtopics in diversification. For example, query "Olympic" may have an original facet about aquatics "diving, swimming, synchronized swimming, water polo", and another original facet about equestrian "dressage, eventing, jumping". In fact, they both are about sports in Olympics and should be grouped together to represent better subtopics in diversification.

To solve the above problem, we further cluster original facets into independent subtopics. Formally, for two query facets t_1 and t_2, and the sets of documents D_{t_1} and D_{t_1} where t_1 and t_2 are extracted from, we calculate the distance between t_1 and t_2 based on their Jaccard similarity.

$$Distance(t_1, t_2) = \frac{D_{t_1} \cap D_{t_2}}{D_{t_1} \cup D_{t_2}}.$$

We use the WQT (quality threshold with weighted data points) clustering algorithm[11] to group query facets into clusters based on the above distance functions. We use M_{dis} as the maximum diameter. Therefore, if two facets t_1, t_2 are closed in distance, i.e., $Distance(t_1, t_2) < M_{\text{dis}}$, we group them into one facet $t_1 \bigcup t_2$. After the clustering process, similar facets are grouped together to compose a candidate of query subtopics.

Moreover, we remove unimportant facets and select top meaningful ones as subtopics. Recall that each original facet t has a weight to describe its importance W_t[11]. For each clustered facet, we sum up its original components' weights as its clustered weight. We set a threshold M_{imp} as a baseline and remove facet t if it is not important enough, i.e., $W_t < M_{\text{imp}}$. We sort the rest facets by their importance and select top n facets as our faceted subtopics.

As stated, we can obtain faceted subtopics based on query facets extracted from search results for any query. Each subtopic t is in the form of an item list, including words and phrases. We formulate our faceted subtopics and each internal subtopic as: $T = \{t_1,\ t_2,\ ..., t_n\}$ and $t_i = \{t_i^1,\ t_i^2,\ ..., t_i^{|t_i|}\}$. Note that Dang and Croft[10] also indicated a subtopic by a set of terms $t_i = \{t_i^1, t_i^2, ., t_i^{|t_i|}\}$. They separated each t_i^j as an independent subtopic and built subtopics in terms $T = \{t_1^1, t_1^2, ..., t_1^{|t_1|}, ..., t_n^1, t_n^2, ..., t_n^{|t_n|}\}$. In contrast, we group terms together to present one subtopic.

After generating faceted subtopics, we calculate the probability $P(t_i|q)$ that subtopic t_i satisfies query q. This probability is hard to estimate for many intent-aware diversification algorithms as their subtopics are predefined or adopted from other resources like Google suggestions. In contrast, our faceted subtopics are generated from search results and their weights straightly reflect their importance in the query. Hence we define $P(t_i|q)$ as the importance ratio of subtopic t_i in query q:

$$P(t_i|q) = \frac{W_{t_i}}{\sum_{t \in T} W_t}\ . \tag{1}$$

In addition, we evaluate the probability $P(t_i^j|t_i)$ that internal item t_i^j satisfies subtopic t_i. We firstly estimate the importance of t_i^j in t_i. According to [11], the importance of an item depends on how many websites contain the item in their lists and its ranks in these lists. For item t_i^j, its importance is as follows:

$$W_{t_i^j} = \sum_{s \in site(t_i)} \frac{1}{\sqrt{AvgRank(t_i^j, s)}}\ .$$

Here $site(t_i)$ denotes the websites that create subtopic t_i. $AvgRank(t_i^j, s)$ is the average rank of t_i^j within all lists extracted from website s. The above equation shows that an important item should be ranked higher by its creator than less important ones in the original list. Based on the importance distribution of items in a subtopic, we calculate $P(t_i^j|t_i)$ as the importance ratio of item t_i^j in subtopic t_i:

$$P(t_i^j|t_i) = \frac{W_{t_i^j}}{\sum_{t_i^{j'} \in t_i} W_{t_i^{j'}}} .$$

Hence we have two sum-to-one probabilities: $P(t_i|q)$, the importance of subtopic t_i in query q and $P(t_i^j|t_i)$, the importance of item t_i^j in its subtopic t_i. They can be automatically computed based on query facets for any query.

Table 2 shows the details of top five faceted subtopics for query "Olympic". The underlying query facet is easy to understand according to the internal items of the subtopic. For instance, subtopic 1 represents the sports of Olympic Games as all its internal items are sport names, and subtopic 2 shows the related countries of Olympic Games (hosted or attended) as its internal items are country names. Similarly, subtopic 3, subtopic 4 and subtopic 5 indicate the related colors of Olympics, the types of Olympics, and the hosted cities of Olympics respectively. Since a group of items displays the details for the subtopic, the relevance between the document and the subtopic can be estimated by the frequencies of the topic items appeared in the document and the weights of the items. The related calculations will be introduced in the next section.

4 Diversification Algorithms

Based on previously generated subtopics, the problem of diversification can be described as follows: for a given query q, let T be the faceted subtopics, obtained from query facets in advance; $P(t_i|q)$ and $P(t_i^j|t_i)$ denote the importance of subtopic t_i in query q and internal item t_i^j in subtopic t_i, calculated by faceted subtopics; R indicates the initial documents set without diversification, which is ranked by some classical ranking algorithm. Our faceted diversification algorithms use T, $P(t_i|q)$, $P(t_i^j|t_i)$ to select a diversified ranked list S of k documents from R.

In this paper, we adapt traditional diversification frameworks for faceted subtopics and propose faceted diversification models. Specifically, we select three state-of-the-art algorithms, the classic MMR[6] algorithm, the famous IASelect[7] algorithm, and the ACSL[9] algorithm, the diversity task winner of TREC 2009 Web track. For each model, we redefine the core functions of their objectives based on faceted subtopics, and maintain their frameworks basically the same.

4.1 Faceted MMR Model

The faceted MMR model (FMMR) is extended from the classic Maximal Marginal Relevance (MMR) algorithm, proposed by Carbonell and Goldstein[6]. MMR diversifies search results by making a trade-off between relevance and novelty based on the similarity of document content. In FMMR, we use faceted subtopics to redefine these similarity functions.

4.1.1 Objective

The general objective is to select the document d^* that maximizes relevance and minimizes similarity to higher ranked documents. Relevance and novelty are measured by two similarity functions: Sim_1, the relevance similarity between document d and query q, and Sim_2, the document similarity among the selected document set S. Parameter λ controls the degree of the trade-off.

$$d^* = \arg \max_{d \in R \setminus S} (\lambda Sim_1(d, q) - (1 - \lambda) \max_{d' \in S} Sim_2(d, d')).$$

4.1.2 Similarity Functions

As there is no subtopic for either the document or the query in original MMR, diversification is conducted through the choice of similarity functions in FMMR. We redefine the above similarity functions by our faceted subtopics.

First, we present Sim_1, the similarity between document d and query q, according to the initial document rank number.

$$Sim_1(d, q) = \frac{1}{\sqrt{rank(d)}}. \quad (2)$$

Note that $rank(d)$ is the initial rank number of document d in the initial search results of query q from a commercial search engine. This similarity is an easy transformation of the rank, without knowing the details of the ranking algorithm. This transformation makes a soft decline of Sim_1, and produces proper values for the linear combination in (3).

We next calculate Sim_2, the similarity between two documents, by their similarities on the subtopics. We formulate the probabilities of one document and the subtopics as a vector. Given two documents d and d' and their vectors $\boldsymbol{V}_d$ and $\boldsymbol{V}_{d'}$, we compute the cosine distance between $\boldsymbol{V}_d$ and $\boldsymbol{V}_{d'}$, and take the result as the similarity between d and d'.

$$Sim_2(d, d') = \cos\theta(\boldsymbol{V}_d, \boldsymbol{V}_{d'}).$$

Note that $\theta(\boldsymbol{V}_d, \boldsymbol{V}_{d'})$ denotes the angle between two vectors, and $\cos\theta(\boldsymbol{V}_d, \boldsymbol{V}_{d'})$ represents the traditional cosine

similarity of two vectors as follows, where V_d^i is the i-th item of vector $\boldsymbol{V}_d$.

$$cos\theta(\boldsymbol{V}_d, \boldsymbol{V}_{d'}) = \frac{\boldsymbol{V}_d \cdot \boldsymbol{V}_{d'}}{|\boldsymbol{V}_d| \cdot |\boldsymbol{V}_{d'}|} = \frac{\sum_{i=1}^{n} V_d^i \times V_{d'}^i}{\sqrt{\sum_{i=1}^{n}(V_d^i)^2} \times \sqrt{\sum_{i=1}^{n}(V_{d'}^i)^2}}.$$

Let us interpret the subtopic vectors in detail. Recall that $T = \{t_1, t_2, ..., t_n\}$ are faceted subtopics generated from the documents of query q. We formulate the vector $\boldsymbol{V}_d = (P(d|t_1), P(d|t_2), ..., P(d|t_n))$ to describe the probability distribution of the subtopics on document d. And we use $P(d|t_i)$ to denote the probability that d satisfies subtopic t_i.

$$P(d|t_i) = \sum_{t_i^j \in t_i} (C(t_i^j, d) \times P(t_i^j|t_i)) \times P(t_i|q).$$

Recall that t_i^j is an item belonging to subtopic t_i. $C(t_i^j, d)$ is the count that t_i^j appears in document d. $P(t_i^j|t_i)$ is the weight of t_i^j in t_i. Then the product $C(t_i^j, d) \times P(t_i^j|t_i)$ is the probability that d satisfies t_i^j in t_i. For a given subtopic t_i, summing up the product over all items in t_i, multiplied by $P(t_i|q)$, the weight of t_i in query q, gives the probability that d satisfies t_i in q.

Note also that two documents are compared by their subtopic probability distributions, according to the cosine similarity of their vectors. If they share a similar distribution on the subtopics, it is more likely that they are similar documents. Otherwise, these documents are more likely to be different.

4.1.3 Algorithm

We propose the FMMR algorithm in Algorithm 1, which selects S from R to maximize relevance Sim_1 and minimize similarity Sim_2 to higher ranked documents.

Initially, we calculate the subtopic probability distributions for all documents. The algorithm then selects one document at a time. At every step, it chooses the document with the highest combination value of a similarity score with respect to the query considering its initial rank, and a dissimilarity score about the already selected documents by their subtopic vector cosine similarities. This marginal relevance tries to output the document that reduces redundancy and maintains query relevance.

Algorithm 1. FMMR
Input: $k,\ q,\ R,\ \lambda,\ T,\ P(t_i|q),\ P(t_i^j|t_i)$
Output: re-ranked set of documents S

```
S = ∅
for d ∈ R do
  Sim_1(d, q) = 1/√rank(d)
  ∀t_i ∈ T, P(d|t_i) = P(t_i|q) × (Σ_{t_i^j ∈ t_i} C(t_i^j, d) × P(t_i^j|t_i))
  V_d = (P(d|t_1), P(d|t_2), ..., P(d|t_n))
end for
while |S| < k do
  for d ∈ R do
    Sim_2(d, S) = arg max_{d' ∈ S} Sim_2(d, d')
  end for
  d* = arg max_{d ∈ R}(λ × Sim_1(d, q) − (1 − λ) × Sim_2(d, S))
  ∀d ∈ R, Sim_2(d, d*) = cos θ(V_d, V_{d*})
  S = S ∪ {d*}
  R = R\{d*}
end while
return S
```

4.1.4 Complexity

The complexity of FMMR depends on the documents selection iterations. It follows the framework of MMR that compares all the unselected documents with the previously selected documents. Let m=$|R|$ be the total document number, the time complexity is $(m + (m-1) \times 1 + ... + (m-(k-1)) \times (k-1)) = O(mk^2 - k^3)$. It holds that $m \geqslant k$; hence the complexity is actually $O(mk^2)$. Note that, when comparing two documents, FMMR compares their subtopic vectors, while MMR compares all their content. Thus FMMR is more efficient than MMR, which is further confirmed in the experiment.

4.2 Faceted IASelect Model

IASelect is a state-of-the-art diversification algorithm proposed by Agrawal *et al.*[7] It provides an objective function to minimize the risk that top k results all fail to satisfy an user. IASelect explicitly considers the diversity of search results through topical categories. The topical categories are predefined based on an open directory project (ODP) taxonomy④, to classify queries and documents. They are summarized in words or phrases as independent subtopics, which is not a good representation as previously discussed.

We propose a faceted IASelect model (FIASelect) by using faceted subtopics instead of categories. We redefine two relevance functions: the probability that a subtopic belongs to a query, and the probability that a document is relevant to a subtopic. We apply our

④ODP. http://www.dmoz.org/, May 2015.

relevance functions to the objective function, and implement it by a greedy algorithm.

4.2.1 Objective

The objective is to maximize the probability that at least one document in the top k results is useful for the average user. Given a query q, a set of documents R, a probability distribution of subtopics for the query $P(t_i|q)$, the relevance values of the documents $V(d|q,t_i)$, and an integer k, the objective selects documents S from R with maximal score calculated by the following objective function:

$$P(S|q) = \sum_{t_i \in T} P(t_i|q) \left(1 - \prod_{d \in S} (1 - V(d|q,t_i))\right).$$

Note that the relevance value $V(d|q,t_i)$ denotes the probability that document d satisfies a user that issues query q with the intended subtopic t_i, and the value $(1 - V(d|q,t_i))$ represents the probability that d fails to satisfy t_i in q. Therefore, given a subtopic t_i, its product is the probability that the whole set of documents S fails to satisfy. The value $(1 - \prod(...))$ equals the probability that some document satisfies subtopic t_i. Finally, summing up all the subtopics, weighted by $P(t_i|q)$, gives the probability that the set of documents S satisfies the average user who issues query q.

4.2.2 Relevance Functions

The core functions in the above objective are the probability distribution of subtopics for query $P(t_i|q)$, and the relevance values of documents $V(d|q,t_i)$, which should be carefully obtained to guarantee the optimality and approximation of the algorithm. In practice, Agrawal *et al.*[7] employed the query classification by algorithm[32] to get $P(t_i|q)$. They classified documents by Rocchio classifier[33], and derived $V(d|q,t_i)$ from relevance score obtained by a commercial search engine.

In FIASelect, we estimate $P(t_i|q)$ and $V(d|q,t_i)$ by faceted subtopics. Recall that we already calculate $P(t_i|q)$ as the weight of subtopic t_i with respect to query q by (1) in Section 3. Nextly, we redefine $V(d|q,t_i)$, the relevance value of document d satisfying subtopic t_i in query q. Following IASelect, we assign $V(d|q,t_i)$ by making a trade-off between query relevance $P(d|q)$, the probability that document d satisfies the user that issues query q, and subtopic relevance $P(d|t_i)$, the probability that document d satisfies the user with subtopic t_i. We use parameter λ to control the degree of the trade-off.

$$V(d|q,t_i) = \lambda \times P(d|q) + (1 - \lambda) \times P(d|t_i). \quad (3)$$

Since the search result documents are already ranked by their relevance initially, we set $P(d|q)$ based on the rank position $rank(d)$ of document d, like (2).

$$P(d|q) = \frac{1}{\sqrt{rank(d)}}. \quad (4)$$

According to our faceted subtopics, we estimate subtopic relevance $P(d|t_i)$ by considering all items in the subtopic. For each item, we consider its frequency in the document, and its weight in the subtopic. Thus the document with more important items is viewed as more relevant to the subtopic. Given a subtopic $t_i = \{t_i^1, t_i^2, ..., t_i^{|t_i|}\}$, we formulate $P(d|t_i)$, the relevance between document d and subtopic t_i as:

$$P(d|t_i) = \frac{\sum_{t_i^j \in t_i} C(t_i^j, d) \times P(t_i^j|t_i)}{\max\limits_{d' \in R} \sum_{t_i^j \in t_i} C(t_i^j, d') \times P(t_i^j|t_i)}. \quad (5)$$

In the above equation, the value $C(t_i^j, d)$ denotes the frequency of subtopic item t_i^j in document d. Summing up all items in subtopic t_i, weighted by $P(t_i^j|t_i)$, the weight of item t_i^j in subtopic t_i, gives the unnormalized subtopic relevance of document d. Then normalizing the value by the maximal unnormalized value of all documents R, gives $P(d|t_i)$, the probability that document d satisfies subtopic t_i.

Note that $P(d|q) \leqslant 1$, $P(d|t_i) \leqslant 1$, and $\lambda \in [0,1]$. It is clear that parameter λ is the key to balance the query relevance and the subtopic relevance. When $\lambda = 0$, the value $V(d|q,t_i)$ is only about query q. When $\lambda = 1$, the value $V(d|q,t_i)$ totally depends on subtopic t_i. We analyze the influence of λ in Section 5.

4.2.3 Algorithm

A greedy algorithm is proposed for FIASelect in Algorithm 2, which records S to maximize the objective in Subsection 4.2.1.

Note that $U(t_i|q,S)$ denotes the conditional probability that subtopic t_i belongs to query q when every document in set S fails to satisfy the user. To begin with, before any document is selected, we set $U(t_i|q,\emptyset) = P(t_i|q)$, and calculate two values for all documents: query relevance $P(d|q)$ and subtopic relevance $P(d|t_i)$.

Algorithm 2. FIASelect
Input: $k,\ q,\ R,\ \lambda,\ T,\ P(t_i|q),\ P(t_i^j|t_i)$
Output: re-ranked set of documents S

$S = \emptyset$
$\forall t_i \in T,\ U(t_i|q,S) = P(t_i|q)$
for $d \in R$ **do**
  $P(d|q) = \frac{1}{\sqrt{rank_d}}$
  $\forall t_i \in T,\ P(d|t_i) = \frac{\sum_{t_i^j \in t_i} C(t_i^j,d) \times P(t_i^j|t_i)}{\arg\max_{d' \in R} \sum_{t_i^j \in t_i} C(t_i^j,d') \times P(t_i^j|t_i)}$
  $\forall t_i \in T,\ V(d|q,t_i) = \lambda \times P(d|q) + (1-\lambda) \times P(d|t_i)$
end for
while $|S| < k$ **do**
  for $d \in R$ **do**
    $g(d|q,T,S) = \sum_{t_i \in T}(U(t_i|q,S) \times V(d|q,t_i))$
  end for
  $d^* = \arg\max_{d \in R}\ g(d|q,T,S)$
  $S = S \cup \{d^*\}$
  $\forall t_i \in T,\ U(t_i|q,S) = 1 - V(d^*|q,t_i) \times U(t_i|q,S \backslash \{d^*\})$
  $R = R \backslash \{d^*\}$
end while
return S

Then we select one document at a time. At each iteration, we choose the document that has the largest marginal value, $g(d|q,T,S)$, computed as the product of the conditional probability of the subtopic, $U(t_i|q,S)$, and the relevant probability of the document and the subtopic, $V(d|q,t_i)$. This marginal value can be interpreted as the probability that the selected document satisfies the user when all previously selected documents fail to satisfy. Note that the conditional distribution will be updated at the end of the loop, to reflect the inclusion of the new document to the selected result set.

4.2.4 Complexity

The complexity of FIASelect depends on the document selection iterations, which selects the document with the maximum target value. The time complexity is $(m + (m-1) + ... + (m-(k-1))) = O(mk)$. Compared with IASelect, FIASelect does a little extra work at computing $V(d|q,t_i)$, but their total time costs are basically the same (see Subsection 5.5).

4.3 Faceted ACSL Model

ACSL is a multi-dimensional diversification algorithm, proposed by Dou *et al.*[9] It won the first place in the diversity task of TREC 2009 Web track. ACSL provides a framework to combine different types of subtopics from different data resources to promote diversity. It implements four types of subtopics, including anchor texts, query logs, search result clusters, and web sites.

In this paper, we propose the FACSL model, by integrating our faceted subtopics as a new data source into the original ACSL algorithm. This combination is workable because the original framework of ACSL is open to incorporate other independent types of subtopics. For the original four types of subtopics, we implement them in the same way as ACSL. For our faceted subtopics, we redefine their relevance functions to fit the basic framework.

4.3.1 Objective

We implement FACSL based on the topic richness model[9], which aims to cover as many subtopics as possible in various data sources, and maintain high relevance to query.

$$d^* = \arg\max_{d \in R \backslash S} (\lambda \times P(d|q) + (1-\lambda) \times \sum_{\mathcal{T} \in \mathbb{T}} P(\mathcal{T}|\mathbb{T}) \times v(d,S,\mathcal{T})),$$

where $P(d|q)$ is the relevance of d in q, and we calculate it by $1/\sqrt{rank(d)}$ as (4). $\mathbb{T}$ denotes different types of subtopics and $P(\mathcal{T}|\mathbb{T})$ shows the weight of subtopic type $\mathcal{T}$ in all types $\mathbb{T}$. We set each subtopic type with uniform weight $P(\mathcal{T}|\mathbb{T}) = \frac{1}{|\mathbb{T}|}$ in the experiment. $v(d,S,\mathcal{T})$ is the relevant score of document d in terms of subtopic $\mathcal{T}$ under the condition of selected documents S. Parameter λ is used to trade off between query relevance and subtopic coverage.

4.3.2 Relevance Function

For the original subtopic types, including anchor texts, query logs, search result clusters, and web site, we calculate their $v(d,S,\mathcal{T})$ exactly the same as the original ACSL algorithm. For the new subtopic type, faceted subtopics, we redefine the relevance functions based on the basic framework and compute its relevant score $v(d,S,T)$ as follows.

$$v(d,S,T) = \sum_{t_i \in T} P(t_i|T) \times \phi(t_i,S) \times P(d|t_i).$$

Here $P(t_i|T)$ is the importance of subtopic t_i in subtopic category T, and $P(d|t_i)$ is the relevance of document d and subtopic t_i. $\phi(t_i,S)$ is the decayed weight of subtopic t_i when document set S has been already selected. Given subtopic t_i, assuming that all documents in S are independent, we utilize the following function to derive $\phi(t_i,S)$:

$$\phi(t_i,S) = \prod_{d \in S}(1 - P(d|t_i)).$$

To be fair for comparisons between algorithms in later experiments, we calculate $P(t_i|T)$, $P(d|q)$, and $P(d|t_i)$ by the same relevance functions as previously introduced. Hence we have $P(t_i|T) = P(t_i|q)$ predefined in Section 3, $P(d|q) = 1/\sqrt{rank(d)}$ as (4), and $P(d|t_i) = \sum_{t_i^j \in t_i} C(t_i^j, d) \times P(t_i^j|t_i) / \max_{d' \in R} \sum_{t_i^j \in t_i} C(t_i^j, d') \times P(t_i^j|t_i)$ as (5).

4.3.3 Complexity

The complexity of FIASelect depends on the objective function, which selects the best k documents from initial ranked m documents, while considering the overlap of selected documents. The time complexity is $O(mk)$. Compared with ACSL, FACSL costs more time to integrate more subtopics, but the difference is not big because they share the same time complexity of the framework (see Subsection 5.5).

5 Experimental Results

We conduct an extensive study to understand the effectiveness and stability of the faceted subtopics and their algorithms presented in the paper. In the experiment, we aim to answer a main question: can we improve the diversification performance with our faceted subtopics?

5.1 Setup

Datasets. We use the public data collection ClueWeb09 in our experiment. The collection consists of one billion web pages in ten languages, collected in January and February of 2009. We use the query set of TREC 2009 Web track[5]. It includes 50 queries, each of which has three to eight manually edited subtopics. There are 243 subtopics in total, and 199 of them have at least one judged relevant document. To the best of our knowledge, it is the first public query set with explicit diversity relevance judgments. We retrieve the top 1 000 results for each query.

Baseline Models. As our proposed search result diversification models (in Section 4) need an initial set of search results, we implement the MSRA2000 model[34] as our baseline ranking function (Baseline). In the adhoc task of TREC 2009 Web track, this ranking function (named MSRANORM) generated reasonably good results[5].

We implement seven classic or state-of-the-art diversification algorithms as our baseline models: MMR, IASelect, ACSL, xQuAD, PM2, $\text{xQuAD}_{\text{term}}$, and PM2_{term}. We have already introduced MMR, IASelect and ACSL in Section 4 and recall that our faceted diversification algorithms are extended from them. Specifically, for MMR[6], we generate the TF-IDF term vector based on document content, and use the cosine similarity of the TF-IDF vectors to measure the similarity between two documents. For IASelect[7], we predefine 16 subtopic categories, and classify queries and documents by Shen *et al.*'s classifiers[35]. For ACSL[9], we employ the subtopic mining methods in the same data sources, and we select the topic richness model as the ACSL algorithm. xQuAD and PM2 are two state-of-the-art diversification algorithms. xQuAD[8] iteratively selects the document to cover the most subtopics which are uncovered by previously selected documents. PM2[22] finds the best unsatisfied subtopic by previously selected documents, and chooses the best document by the selected subtopic. They both use Google Suggestions as their subtopics. $\text{xQuAD}_{\text{term}}$, and PM2_{term}[10] are term level diversification algorithms extended from xQuAD and PM2. They split key terms from original subtopics and use them as subtopics. We cut Google Suggestions into terms by Stanford Tokenizor⑤. Note that we do not implement the DSPApprox method introduced by [10] as $\text{xQuAD}_{\text{term}}$ and PM2_{term} outperformed it in [10].

Evaluation Metrics. The evaluation metrics used in the diversity tasks of TREC2009 Web Track are adopted, including normalized discounted cumulative gain (α-NDCG)[14] and intent-aware precision (Precision-IA)[7]. For α-NDCG, the discounted gain of a document depends on how much novel information it provides. For Precision-IA, the precision value is the average precision values of all intents. Their default parameter settings are used according to TREC2009 Diversity task. We report α-NDCG and Precision-IA at retrieval depths 5, 10, and 20 respectively. We use two-tailed t-test for statistical significance and report significant differences when $p \leqslant 0.05$.

5.2 Overall Results

Table 3 shows the evaluation results of the baseline models, i.e., MMR, PM2, PM2_{term}, IASelect, xQuAD, $\text{xQuAD}_{\text{term}}$, and ACSL, and the adaptive faceted models, i.e., FMMR, FIASelect and FACSL, in terms of α-NDCG and Precision-IA (P-IA).

⑤Stanford Tokenizor. http://nlp.stanford.edu/software/tokenizer.shtml, May 2015.

Table 3. Performance Comparison on TREC 2009

	α-NDCG@5	α-NDCG@10	α-NDCG@20	Precision-IA@5	Precision-IA@10	Precision-IA@20
Baseline	0.243 6	0.285 7	0.327 9	0.115 9	0.104 6	0.098 2
MMR*	0.225 2	0.269 4	0.305 5	0.093 3	0.089 4	0.079 6
PM2°	0.263 6	0.283 9	0.330 6	0.127 2	0.106 3	0.094 3
PM2$^{\bullet}_{\text{term}}$	0.238 0	0.276 2	0.310 1	0.101 0	0.091 4	0.077 1
IASelect*	0.267 6	0.306 3	0.342 7	0.118 7	0.115 1	0.098 2
xQuAD$^{\divideontimes}$	0.293 6	0.319 4	0.353 8	0.136 1	0.113 6	0.100 4
xQuAD$^{\diamond}_{\text{term}}$	0.288 3	0.317 9	0.355 7	0.125 1	0.106 2	0.096 1
ACLS†	0.291 3	0.324 2	0.374 0	0.131 4	0.115 1	**0.105 8**
FMMR	0.236 0	0.273 9	0.319 0	0.108 3	0.101 6*	0.098 2*
FIASelect	0.309 6*	0.329 6*	0.361 2*	0.140 8$^{*\bullet}_{\diamond}$	0.120 0$^{*\bullet}$	0.098 2$^{*\bullet}$
FACSL	**0.321 4**$^{*\bullet}$	**0.349 7**$^{*\bullet}_{\circ}$	**0.389 4**$^{*\bullet}_{\circ}$	**0.147 7**$^{*\bullet}_{\divideontimes\star}$	**0.127 5**$^{*\bullet}_{\divideontimes\dagger}$	**0.105 8**$^{*\bullet}_{\divideontimes}$

Note: The best result is in bold. Statistical significant differences between the faceted methods (FMMR, FIASelect, FACSL) and the baseline methods (MMR, PM2, PM2$_{\text{term}}$, IASelect, xQuAD, xQuAD$_{\text{term}}$, and ACSL) are marked with $*, \circ, \bullet, \star, \divideontimes, \diamond, \dagger$ respectively.

Without utilizing subtopics, MMR performs the worst in terms of all metrics, even worse than Baseline. FMMR implicitly considers faceted subtopics in measuring the document similarity. It significantly outperforms MMR in terms of P-IA, but it still underperforms Baseline in terms of most metrics. This indicates that implicitly considering user intents is not good enough in diversification.

As state-of-the-art intent-aware diversification algorithms, PM2, PM2$_{\text{term}}$, IASelect, xQuAD, xQuAD$_{\text{term}}$, and ACSL outperform the non-diversified baseline (Baseline) in terms of all metrics. ACSL outperforms all the other baselines in terms of most metrics as it won the first place at diversity task in TREC 2009 Web track. xQuAD is the second best baseline model which slightly outperforms ACSL in terms of α-NDCG@5 and P-IA@5. xQuAD$_{\text{term}}$ and PM2$_{\text{term}}$ work close to their corresponding models as introduced in [10]. Their performance is affected by their subtopics in term level, which may lose relevance when queries contain phrases. Considering query No.1 "obama family tree" as an example, an original subtopic "obama family tree pictures" is split up as "obama", "family", "tree", and "pictures", which may mislead term level diversification algorithms to view documents about "family", "tree", and "pictures" as relevant.

Our proposed intent-aware faceted diversification models, i.e., FIASelect and FACSL, outperform all baseline methods on the TREC 2009 dataset, including their corresponding models IASelect and ACSL. Both FIASelect and FACSL have statistically significant improvements over the baseline models, in terms of α-NDCG and P-IA ($p < 0.05$ with two-tailed t-tests). This indicates that by leveraging faceted subtopics, our faceted diversification models outperform the state-of-the-art diversification methods. The results clearly show that incorporating faceted intents can improve search result diversification.

Moreover, FACSL has statistically significant improvements over all the baseline models. Compared with the best baseline model ACSL, FACSL has a more-than-three-point gain in terms of α-NDCG and a more-than-one-point gain in terms of P-IA. Recall that ACSL is a multi-dimensional algorithm, and FACSL integrates the faceted subtopics as a new type of data source into the framework. As ACSL itself is a good diversification approach (the best baseline model in Table 3) and uses four kinds of data sources (anchor texts, query logs, search result clusters, and web sites), the results show that our faceted subtopics based on query facets are complementary with the other data sources. By adding faceted subtopics, FACSL can leverage this strong model and further improve it.

5.3 Effect of Document Numbers

Fig.1 shows the results based on different document numbers. As the trends of different approaches are similar, to save space, we report the results of FIASelect and IASelect on α-NDCG@10 and Precision-IA@10 in the rest part of experimental results. We caution that the gap of the x-axis between 50 and 100 is ignored so as to save space without influencing the general trend.

FIASelect consistently outperforms IASelect in terms of all metrics. Furthermore, with the increase of document numbers, the performance of FIASelect decreases less than the performance of IASelect. FIASelect slightly underperforms Baseline when the document number reaches 100. It shows that the faceted

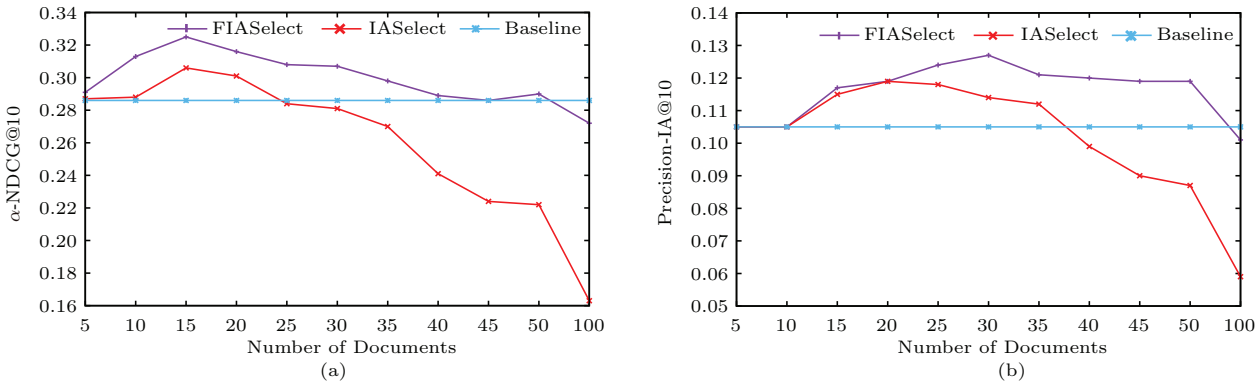


Fig.1. Compare the results of different document numbers for IASelect and FIASelect on (a) α-NDCG@10 and (b) Precision-IA@10.

subtopics have more effective and stable performance than the traditional subtopics.

5.4 Effect of Parameter λ

For the diversification problem, $V(d|q,t_i)$, the probability of document d satisfying subtopic t_i in query q, is an very important function. In FIASelect, we use a linear combination, $V(d|q,t_i) = \lambda \times P(d|q) + (1-\lambda) \times P(d|t_i)$ as (3), to balance the query relevance $P(d|q)$ and the subtopics relevance $P(d|t_i)$. Parameter λ controls the degree of the trade-off, and $\lambda \in [0,1]$. If $\lambda = 0$, documents are selected totally by subtopics; if $\lambda = 1$, documents are chosen by query relevance.

Fig.2 shows three different values of λ on different numbers of documents in FIASelect. Obviously, a larger λ is more stable in all document numbers, as it depends more on query relevance, and a smaller λ has a better performance in most situations, as it cares more about subtopic relevance. This observation matches our expectation and proves that the design of λ is reasonable and effective.

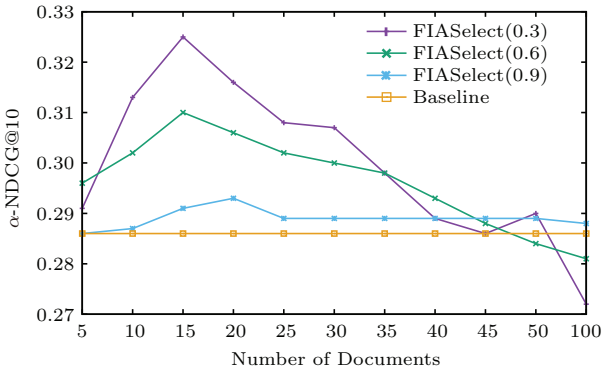


Fig.2. Three λ values of FIASelect. Note that FIASelect(k) indicates the result of FIASelect with $\lambda = k$.

Fig.3 shows the performance of FIASelect when using different λ for different numbers of documents. We run FIASelect on different document numbers for these λ values and denote the point line according to its document number (e.g., Top5 indicates the result of FIASelect on the top 5 documents).

Fig.3(a) shows that, in terms of α-NDCG, as the document number increases, the influence of λ decreases, except for the Top5 documents. The reason is that more irrelevant documents are retrieved and they may hurt the effectiveness of λ. In Fig.3(b), different document numbers represent little impact of λ in terms of Precision-IA. The Top5 overlaps Baseline, as Precision-IA ignores the ranks of documents. Here the effect of λ on different document numbers is less obvious.

We observe that for most document numbers, the results of α-NDCG and Precision-IA change slightly when $\lambda \leqslant 0.4$, and decrease quickly when $\lambda > 0.4$. Thus a better choice for λ is 0.3 or 0.4. We set $\lambda = 0.3$ in our FIASelect algorithm. FACSL uses the same value as it shares a similar equation with FIASelect.

In addition, parameter λ is similarly used in other baseline algorithms, including xQuAD, $\text{xQuAD}_{\text{term}}$, and ACSL, to balance query relevance and subtopic relevance. Fig.4 compares the performance of some baseline models with FIASelect on different settings of λ. It shows that our FIASelect model consistently outperforms all the other algorithms on different settings of λ in terms of α-NDCG and Precision-IA. To be fair, we assign the best λ setting for each baseline model respectively, and report their best results in Table 3. As we introduced before, FIASelect and FACSL still outperform all the baseline models in terms of most metrics.

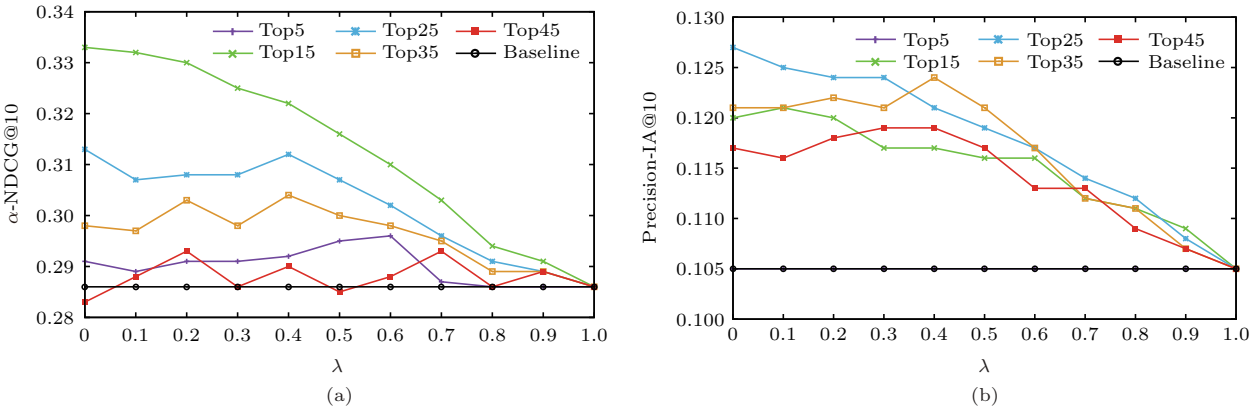


Fig.3. FIASelect results for different λ values on different document numbers. Note that Topk indicates the result of FIASelect on the top k documents. (a) α-NDCG results. (b) Precision-IA results.

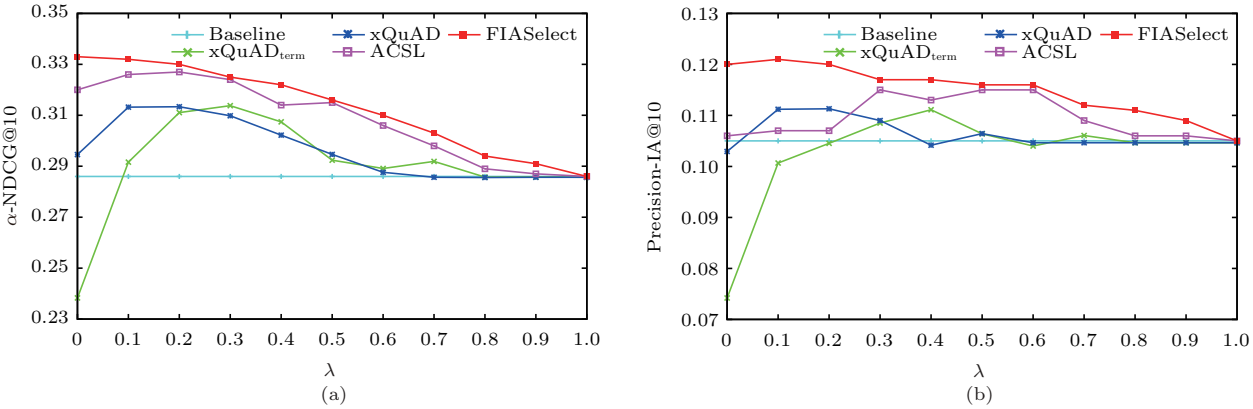


Fig.4. Performance comparison of the best baseline methods and the best faceted methods on different λ values. (a) α-NDCG results. (b) Precision-IA results.

5.5 Performance

Fig.5 compares the average time cost to diversify search results for one query. Here we only analyze the core time of the algorithms, and omit the time cost of data preprocessing (e.g., document loading, word stem), because it costs the same for all diversity algorithms.

It can be seen that, by using traditional subtopics, PM2, PM2$_{\text{term}}$, xQuAD, xQuAD$_{\text{term}}$, and IASelect are very fast. FIASelect is highly close to IASelect, which shows that faceted subtopics do not cost more time than traditional subtopics.

ACSL needs more time as it combines four types of subtopics instead of one type. FACSL is a little slower than ACSL, by dealing with the extra fifth (faceted) subtopics.

MMR gets document similarity by comparing their full content, and thus it is the slowest one, unsurprisingly. FMMR represents document similarity by faceted subtopics, which largely improve the efficiency. When utilizing faceted subtopics, FMMR costs much more time than FIASelect and FACSL, because FMMR's time complexity $O(mk^2)$ is higher than FIASelect and FACSL's time complexity $O(mk)$.

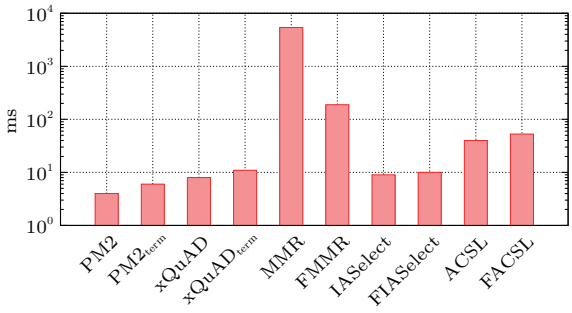


Fig.5. Average running time comparison of baseline methods and faceted methods.

6 Conclusions

In this paper, we showed that the common user intents of a query can have a more complex structure than a group of words or phrases. We studied the problem of mining and utilizing query facets as subtopics. In particular, we extracted query facets from search results for a given query, merged related facets into one cluster, and selected top ranked clusters as the final subtopics. In contrast to common user intents, each of which is a word or phrase, our subtopics are organized as query facets, each of which is a group of words or phrases extracted from search results, and the internal items of a subtopic explain an underlying user intent. The subtopics can be mined for all queries, including rare queries and new queries.

We adapted three state-of-the-art algorithms, i.e., MMR, IASelect, and ACSL, and proposed three corresponding models, i.e., FMMR, FIASelect, and FACSL, to diversify search results based on faceted subtopics. For each model, we redefined the critical functions about the subtopics, and kept the framework settings basically the same with its original model. Experimental results on TREC 2009 Web track data showed that by using faceted subtopics, the adaptive algorithms obtained improved diversity compared with their original models. The FACSL model outperforms state-of-the-art diversity algorithms discussed in this paper. The results indicated that exploiting faceted subtopics from query facets can benefit search result diversification.

References

[1] Jansen B J, Spink A, Saracevic T. Real life, real users, and real needs: A study and analysis of user queries on the web. *Inf. Process. Manage.*, 2000, 36(2): 207–227.

[2] Silverstein C, Marais H, Henzinger M, Moricz M. Analysis of a very large web search engine query log. *SIGIR Forum*, 1999, 33(1): 6–12.

[3] Dou Z, Song R, Wen J R. A large-scale evaluation and analysis of personalized search strategies. In *Proc. the 16th WWW*, May 2007, pp.581–590.

[4] Rafiei D, Bharat K, Shukla A. Diversifying web search results. In *Proc. the 19th WWW*, April 2010, pp.781–790.

[5] Clarke C L A, Craswell N, Soboroff I. Overview of the TREC 2009 web track. In *Proc. the 18th TREC*, November 2009.

[6] Carbonell J, Goldstein J. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proc. the 21st SIGIR*, August 1998, pp.335–336.

[7] Agrawal R, Gollapudi S, Halverson A, Ieong S. Diversifying search results. In *Proc. the 2nd WSDM*, February 2009, pp.5–14.

[8] Santos R L T, Macdonald C, Ounis I. Exploiting query reformulations for web search result diversification. In *Proc. the 19th WWW*, April 2010, pp.881-890.

[9] Dou Z, Hu S, Chen K, Song R, Wen J R. Multi-dimensional search result diversification. In *Proc. the 4th WSDM*, February 2011, pp.475–484.

[10] Dang V, Croft W B. Term level search result diversification. In *Proc. the 36th SIGIR*, July 28–August 1, 2013, pp.603–612.

[11] Dou Z, Hu S, Luo Y, Song R, Wen J R. Finding dimensions for queries. In *Proc. the 20th CIKM*, October 2011, pp.1311–1320.

[12] Kong W, Allan J. Extracting query facets from search results. In *Proc. the 36th SIGIR*, July 28–August 1, 2013, pp.93–102.

[13] Kong W, Allan J. Extending faceted search to the general web. In *Proc. the 23rd CIKM*, Nov. 2014, pp.839–848.

[14] Clarke C L, Kolla M, Cormack G V, Vechtomova O, Ashkan A, Büttcher S, MacKinnon I. Novelty and diversity in information retrieval evaluation. In *Proc. the 31st SIGIR*, July 2008, pp.659–666.

[15] Zhai C, Lafferty J. A risk minimization framework for information retrieval. *Inf. Process. Manage.*, 2006, 42(1): 31–55.

[16] Chen H, Karger D R. Less is more: Probabilistic models for retrieving fewer relevant documents. In *Proc. the 29th SIGIR*, August 2006, pp.429–436.

[17] Zhang B, Li H, Liu Y, Ji L, Xi W, Fan W, Chen Z, Ma W Y. Improving web search results using affinity graph. In *Proc. the 28th SIGIR*, August 2005, pp.504–511.

[18] Santos R L, Macdonald C, Ounis I. Selectively diversifying web search results. In *Proc. the 19th CIKM*, October 2010, pp.1179–1188.

[19] Santos R L, Macdonald C, Ounis I. Intent-aware search result diversification. In *Proc. the 34th SIGIR*, July 2011, pp.595–604.

[20] Yue Y, Joachims T. Predicting diverse subsets using structural SVMs. In *Proc. the 25th ICML*, July 2008, pp.1224–1231.

[21] Radlinski F, Kleinberg R, Joachims T. Learning diverse rankings with multi-armed bandits. In *Proc. the 25th ICML*, July 2008, pp.784–791.

[22] Dang V, Croft W B. Diversity by proportionality: An election-based approach to search result diversification. In *Proc. the 35th SIGIR*, August 2012, pp.65–74.

[23] He J, Hollink V, de Vries A. Combining implicit and explicit topic representations for result diversification. In *Proc. the 35th SIGIR*, August 2012, pp.851–860.

[24] Zhu Y, Lan Y, Guo J, Cheng X, Niu S. Learning for search result diversification. In *Proc. the 37th SIGIR*, July 2014, pp.293–302.

[25] Yu H T, Ren F. Search result diversification via filling up multiple knapsacks. In *Proc. the 23rd CIKM*, November 2014, pp.609–618.

[26] Liang S, Ren Z, de Rijke M. Fusion helps diversification. In *Proc. the 37th SIGIR*, July 2014, pp.303–312.

[27] Lawrie D, Croft W B, Rosenberg A. Finding topic words for hierarchical summarization. In *Proc. the 24th SIGIR*, September 2001, pp.349–357.

[28] Hu Y, Qian Y, Li H, Jiang D, Pei J, Zheng Q. Mining query subtopics from search log data. In *Proc. the 35th SIGIR*, August 2012, pp.305–314.

[29] Abbassi Z, Mirrokni V S, Thakur M. Diversity maximization under matroid constraints. In *Proc. the 19th SIGKDD*, August 2013, pp.32–40.

[30] Bache K, Newman D, Smyth P. Text-based measures of document diversity. In *Proc. the 19th SIGKDD*, August 2013, pp.23–31.

[31] Jameel S, Lam W. An unsupervised topic segmentation model incorporating word order. In *Proc. the 36th SIGIR*, July 28–August 1, 2013, pp.203–212.

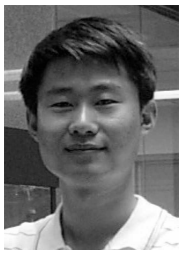
[32] Fuxman A, Tsaparas P, Achan K, Agrawal R. Using the wisdom of the crowds for keyword generation. In *Proc. the 17th WWW*, April 2008, pp.61–70.

[33] Manning C D, Raghavan P, Schütze H. Introduction to Information Retrieval (1st edition). Cambridge University Press, 2008.

[34] Song R, Wen J R, Shi S, Xin G, Liu T Y, Qin T, Zheng X, Zhang J, Xue G R, Ma W Y. Microsoft research Asia at web track and terabyte track of TREC 2004. In *Proc. the 13th TREC*, November 2004.

[35] Shen D, Pan R, Sun J T, Pan J J, Wu K, Yin J, Yang Q. Q^2C@UST: Our winning solution to query classification in KDDCUP 2005. *SIGKDD Explorations*, 2005, 7(2): 100–110.

Sha Hu is a Ph.D. student of computer science at Renmin University of China, Beijing. She received her Bachelor's degree in computer science from Renmin University of China in 2008. She worked at Microsoft Research Asia as a research intern in the Web Search and Mining Group from 2008 to 2013. Her research focuses on information retrieval and Web data extraction.



Zhi-Cheng Dou is an associate professor in the School of Information and Key Laboratory of Data Engineering and Knowledge Engineering, Renmin University of China, Beijing. He received his Ph.D. and B.S. degrees in computer science and technology from the Nankai University, Tianjin, in 2008 and 2003, respectively. After getting his Ph.D. degree, he worked at Microsoft Research as a researcher from July 2008 to September 2014. His research interests include information retrieval, data mining, and big data analytics.

Xiao-Jie Wang is a junior student in the School of information at Renmin University of China, Beijing. His research interest is in the field of information retrieval and he has done some work about search result diversification.

Ji-Rong Wen is a professor at Renmin University of China, Beijing. He is also a National "1000 Talents Project" expert of China. His main research interests include big data management & analytics, information retrieval, data mining and machine learning. He was a senior researcher at Microsoft Research Asia (MSRA) and has filed more than 50 U.S. patents in Web search and related areas. He has published extensively on prestigious international conferences and journals. He is currently the associate editor of ACM Transactions on Information Systems (TOIS). He is a senior member of CCF, ACM and IEEE.