Leveraging Dynamic Query Subtopics for Time-aware Search Result Diversification*

Tu Ngoc Nguyen and Nattiya Kanhabua

L3S Research Center / Leibniz Universität Hannover
Appelstrasse 9a, Hannover 30167, Germany
{tunguyen, kanhabua}@L3S.de

Abstract. Search result diversification is a common technique for tackling the problem of ambiguous and multi-faceted queries by maximizing query aspects or subtopics in a result list. In some special cases, subtopics associated to such queries can be temporally ambiguous, for instance, the query US Open is more likely to be targeting the tennis open in September, and the golf tournament in June. More precisely, users' search intent can be identified by the popularity of a subtopic with respect to the time where the query is issued. In this paper, we study search result diversification for time-sensitive queries, where the temporal dynamics of query subtopics are explicitly determined and modeled into result diversification. Unlike aforementioned work that, in general, considered only static subtopics, we leverage dynamic subtopics by analyzing two data sources (i.e., query logs and a document collection). By using these data sources, it provides the insights from different perspectives of how query subtopics change over time. Moreover, we propose novel time-aware diversification methods that leverage the identified dynamic subtopics. A key idea is to re-rank search results based on the freshness and popularity of subtopics. To this end, our experimental results show that the proposed methods can significantly improve the diversity and relevance effectiveness for time-sensitive queries in comparison with state-of-the-art methods.

1 Introduction

A significant fraction of web search queries are ambiguous, or contain multiple aspects or subtopics [7]. For example, the query apple can refer to a kind of fruit or a company selling computer products. Moreover, the underlying aspects of the query apple inc can be a new Apple product, software updates or it latest press releases. While it is difficult to identify user's search intent for multi-faceted queries, it is common to present results with a high coverage of relevant aspects. This problem has been well studied in aforementioned work on search result diversification [1,5,6,10,15,16]. However, previous work only consider a set of static subtopics without taking into account the temporal dynamics of query subtopics.

* This work was partially funded by the European Commission FP7 under grant agreement No.600826 for the ForgetIT project (2013-2016)

In this paper, we study the search result diversification of *temporally ambiguous, multi-faceted queries*, where the relevance of query subtopics is highly time-dependent. For example, when issuing the query kentucky derby in April, relevant aspects are likely to be about "festival" or "food" referring to the Kentucky Derby Festival, which occurs two weeks before the stakes race. However, at the end of May, other facets like "result" and "winner" should be more relevant to than pre-event aspect. Identifying dynamic subtopics for temporally ambiguous, multi-faceted queries is essential for time-aware search result diversification. In order that, we explicitly extract dynamic subtopics and leverage them into diversifying retrieved results. To the best of our knowledge, none of the aforementioned works considers the temporal changes in query subtopics before.

Our contributions in this paper are as follows. We study the temporal dynamics of subtopics for queries, which are *temporally ambiguous* or *multi-faceted*. We analyze the temporal variability of query subtopics by applying subtopic mining techniques at different time periods. In addition, our analysis results reveal that the popularity of query aspects changes over time, which is possibly the influence of a real-world event. The analysis study is based on two data sources, namely, query logs and a temporal document collection, where time information is available. To this end, we propose different time-aware search result diversification methods, which leverage dynamic subtopics and show the performance improvement over the existing non time-aware methods.

2 Dynamic Subtopic Mining

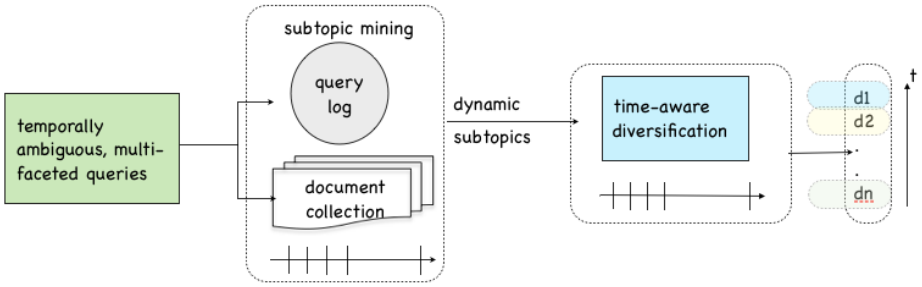


Fig. 1: Pipeline for dynamic subtopic mining and time-aware diversification

In this section, we present our methodology in modeling and mining temporal subtopics from two different datasets. The mined subtopics are input for our time-aware diversification approach. Figure 1 depicts the pipeline process of this paper.

2.1 Mining Subtopics from Query Logs

In our work, we followed a state-of-the-art finding related queries technique proposed in [9]. We applied Markov random walk with restart (RWR) on the

weighted bipartite graph composed of two sets of nodes, namely, queries and URLs. The bipartite graph is constructed using the history information with regards to different time points. Our model for dynamic subtopic mining assigns each subtopic a *temporal weight* that reflects the probability of the relevance of a subtopic at the particular time.

Models In this work, we use real-world Web search logs composed of a set of queries Q, a set of URLs D_{url} and click-through information S. Each query $q \in Q$ is made up of query terms, the hitting time q_t (when a user issues the query) and a set of subtopic C. A clicked URL $u_q \in D_{url}$ refers to a web document returned as an answer for a given query q, which has been clicked by the user. The click-through information composes of a query q, a clicked URL u_q, the position on result page, and its timestamps. A weighted directed bipartite graph $G = (Q \cup D_{url}, E)$, where edges E represent the click-through information from a query $q \in Q$ to a URL $u \in D$. Edges are weighted using click frequency - inverse query frequency (CF-IQF) model. CF-IQF compensates for common clicks on less frequent but distinguished URLs over common clicks on frequent URLs. A subtopic $c \in C$ is represented as a bag of words and associates to a temporal weight $w(c)$, where w is a weighting function.

Clustering subtopic candidates Random walk with restart on the click-through graph provides us a set of related queries. However, these related queries can be duplicated or near-duplicated in their semantics. To achieve finer-grained query subtopics at hitting time q_t, we cluster the acquired queries in a similar approach proposed in [17]. The steps are as follows: (1) Construct a query similarity matrix (using lexical, click and semantic similarity), (2) Cluster related queries (using Affinity Propagation technique), and (3) Extract dynamic query subtopics. Due to the limited space in this paper, readers can refer to [17] for detailed description of the steps.

Temporal subtopic weight We calculate the subtopic weight from query log $w_{query\_log}(c)$ of a subtopic c to a query q solely based on the relatedness score from performing the RWR. For each query cluster $\mathbb{C}_i$ that represents a subtopic c_i, the weight of c, $w(c)$ is the proportion between the total RWR score of all queries in $\mathbb{C}_i$ and of all related queries.

2.2 Mining Subtopics from a Temporal Document Collection

In this section, we make use of Latent Dirichlet Allocation (LDA) [4], an unsupervised method to mine and model latent query subtopics from a relevant set of documents D. Relevant sets of documents are captured at fixed time periods in order to measure the variance of the mined latent subtopics over time. Here, a subtopic $c \in C$ is modeled as multinomial distribution of words, a document $d \in D$ composes of a mixture of topics.

Estimating number of subtopics Deciding the optimum number of subtopics is an important task for assessing the overall query subtopic dynamics. The number of subtopics is expected to change when mining it at different time points. In this work, we follow the approach that proposed by Arun et al. [2] to identify the number of latent subtopics that are naturally present in each partition. The

non-optimum number of subtopics produces the high divergence between the salient distributions derived from two matrix factors (compose of topics-words and documents-topics). In our case, we set the number of topics in a pre-defined range from γ to δ, the chosen number of topic is the one with the minimum KL-divergence value.

Temporal subtopic weight We estimate the weight of a mined subtopic at every hitting time. The weight $w_{docs}(c)$ of a subtopic c reflects the probability that a given query q implies the subtopic c. The temporal distribution that specifies the probability that a given query belongs to a subtopic c, $Pr(c|q)$ derives from the popularity of the subtopic in the studied time slice of the document collection. It is calculated as the proportion between the total probabilities of all documents belongs to a subtopic $Pr(c|d)$ and the number of documents in the time slice. $Pr(c|d)$ is calculated from the Dirichlet prior topic distribution of LDA.

3 Time-aware diversification

Most of the existing diversification approaches mentioned in this paper deploy a greedy approximation approach. We examine three state-of-the-art diversification models (i.e., IA-Select [1], xQuaD [16] and topic-richness [10]). We aim to maximize the utility of the models by fostering recent documents in the ranking, with the assumption that the recency level of a subtopic is linearly proportional to its temporal popularity.

temp-IA-Select The objective function of IA-Select can be expressed using a probabilistic model as:

$$f_s(d) = \sum_c Pr(q|d)Pr(d|c)Pr(c|q) \prod_{d\prime \in S} (1 - Pr(q|d\prime)Pr(d\prime|c)) \quad (1)$$

where S is the selected set of diversified documents from the original result set. Our assumption is our temporal mined subtopics are fresh subtopics and the subtopics tend to favor recent documents. We propose an exponential distribution on the probability of documents $Pr(d)$ with regards to a subtopic c. The document-subtopic probability $Pr(d|c)$ at time t_d, defined as $Pr_{t_d}(d|c)$ is calculated in Equation 2.

$$Pr_{t_d}(d|c) = Pr(c|d)Pr(d|t_d) = Pr(c|d) \cdot \lambda \cdot e^{-\lambda \cdot t_d} \quad (2)$$

We apply $Pr_{t_d}(d|c)$ into the probabilistic objective function of IA-Select to achieve our time-aware objective function (temp-IA-Select), described in Equation 3. With this approach, a document d which is published closer to the hitting time t_q, in essence, has a shorter age t_d will be weighted higher than the one with the same $Pr(c|d)$. Note that for this setting, we do not account for time to calculate the document-query probability, $Pr(d|q)$, that remains unchanged over time. Our intuition is to leverage only exponential distribution of a document d towards certain subtopic d in favoring recent documents in the task of diversifying search

Table 1: *Temporally ambiguous*, multi-faceted AOL queries

harry potter	apple	ncaa	clinton	selection	civil war
mlb	final four	easter	election	march madness	obama
oscar	kentucky derby	nasdaq	nfl	cannes	hurricane
mccain	olympics	opening day	kentucky	tiger	presidential election
triple crown	iraq	preakness	us open	euro 2008	spain

results (according to the mined subtopics C).

$$f_s(d) = \sum_c Pr(c|q)Pr(q|d)Pr(c|d) \cdot \lambda \cdot e^{-\lambda \cdot t_d} \prod_{d\prime \in S} (1 - Pr(q|d\prime)Pr(c|d\prime) \cdot \lambda \cdot e^{-\lambda \cdot t_{d\prime}}) \tag{3}$$

temp-xQuaD Analogously, we modified the probabilistic model of xQuaD. Different from IA-Select, xQuaD introduces the parameter α, to control the trade-off between relevance and diversity. The objective function of temp-xQuaD is given in Equation 4.

$$f_s(d) = (1-\alpha)Pr(d|q) + \alpha \sum_c Pr(c|q)Pr(c|d) \cdot \lambda \cdot e^{-\lambda \cdot t_d} \prod_{d\prime \in S} (1 - Pr(c|d\prime) \cdot \lambda \cdot e^{-\lambda \cdot t_{d\prime}}) \tag{4}$$

temp-topic-richness Differently, the subtopics in topic-richness are modeled as a set of different data sources. The objective function of topic richness model is the generalization of IA-Select and xQuaD framework. Hence, we inject the temporal factor into the model analogously to what we did with temp-IA-Select and temp-xQuaD.

4 Experiments

In this section, we first investigate the quality of the subtopic mining from multiple sources. We then evaluate the performance of our time-aware diversification models on top of the mined subtopics on different metrics.

4.1 Evaluating Subtopics Mined from Query Log

The dataset used is the query log of AOL search engine (from March to May 2006), which has more than 30 million queries and 20 million click information. We time partition the query log into 12 accumulated parts (each approx. one week length), according to 12 simulated fixed hitting times. We manually selected 30 *temporally ambiguous*, multi-faceted queries, as shown in Table 1 for further studies.

Preliminary analysis in the AOL Query Log Due to the short time span (three months) of AOL, we decide to present a small study on the dynamic query aspects at a specific period. Table 2 shows the subtopics of the query ncaa at three different hitting times: *March 14th*, *March 31th* and *April 7th*. We speculate that the change in temporal aspects of a query tends to bind to the appearance of a happening event. We observe that the subtopics retrieved

Table 2: Top-10 query subtopics (ordered by temporal weights) of *ncaa* in AOL

March 14th	March 31th	April 07th
0.0132 · march madness schedule	**0.0100** · oakland raiders	**0.0122** · ncaa women's basketball tournament
0.0117 · ncaa basketball tournament	**0.0090** · ncaaw	**0.0053** · ncaa basketball tournament
0.0068 · nfl draft	**0.0042** · tito francona	**0.0049** · cbs sports line
0.0048 · selection sunday	**0.0031** · ncaa brackets	**0.0033** · ncaaw
0.0037 · oakland raiders	**0.0029** · ncaa division ii	**0.0031** · ncaa final four
0.0032 · 2006 ncaa tournament bracket	**0.0024** · andy goram	**0.0029** · ncaa wrestling
0.0026 · brad hopkins released nfl	**0.0024** · lakers	**0.0028** · march madness bracket
0.0023 · roger clemens	**0.0024** · ncaa women's basketball tournament bracket	**0.0019** · ncaa basketball results
0.0021 · ncaa division ii	**0.0021** · ncaa basketball brackets	**0.0009** · andy goram
0.0014 · college basketball	**0.0021** · nit brackets	**0.0009** · ncaa division ii

Table 3: Statistics on subtopics quality

	Mean	Variance	Kurtosis
Coherence	0.23	0.11	3.402
Distinctness	0.831	0.072	1.031
Plausibility	0.158	0.193	4.227
Completeness	0.654	0.316	-1.029

on *March 14th* reflect the *pre-event* aspects of the event march madness[1], e.g., *2006 ncaa tournament bracket*, *march madness schedule*. When mining subtopics on March 31st, a new subtopic emerges (*ncaa women's basketball tournament bracket*) refers about a new sub-event that occurs later on March 18th. We capture subtopics with post and late-event aspects when mining them on April 7th, i.e., *ncaa basketball results* for *post-event*, *ncaa final four* for *late-event*.

Evaluation metrics As we work on a specific set of queries, there is no standard test collection to evaluate the quality of the subtopic mining. We hence applied a novel assessment proposed by Radlinski *et al.* [14], where they defined four distinct metrics (i.e., coherence, distinctness, plausibility and completeness). Coherence indicates the level of explicitness is a subtopic, distinctness measures the distance between subtopics in terms of the relevant documents (URLs). Plausibility indicates how many percent of users who issue query q are satisfied with subtopic c. while completeness shows how complete is the set of subtopics C. In the query log context, we assume the relevance of a URL as being clicked on.

Experimental results We report the detailed results for coherence, distinctness, plausibility and completeness in Table 3. We empirically report the results of 10 (out of 30) queries that are most active in the AOL time span (by analyzing the time series) the queries based on statistical measurements i.e., mean, variance and kurtosis. All the reported results are in the range from 0 to 1, with 1 as the best result. The high distinctness result indicates the high diversity of our mined subtopics in the query log. The result for completeness is 0.654, meaning our mined subtopics give a good coverage to the possible aspects of a query. The rather low results for coherence and plausibility (β reported at 0.5) are expected because the experiment is conducted based only on the query log. We focus on mining facets of the query and the distance between facets and its original query

[1] Note that march madness is the synonym of the ncaa basketball tournament that is held every March and final four refers to the later stage of the tournament.

can be long. The similarity metric for the measurement is based only on Jaccard Similarity between sets of documents, without looking into the document contents (that are not available), also contributes to the low result.

4.2 Evaluating Subtopics Mined from a Temporal Document Collection

We used the TREC Blogs08 as a temporal document collection to conduct the experiment. The collection is crawled from January 2008 to February 2009, with more than 28 millions web documents in total (more than 70 thousand per day). For our task, the accurate and trustworthy timestamps of documents are necessary. However, timestamps associated to web documents in general are less reliable due to its decentralized nature and there is no standard for time and date [11]. We determined the timestamp of a blog document based on the three sources (ranked by reliability order), namely, document content, document URL and the crawled date. For each query, we construct a different LDA model on its partitioned results (there are 14 according to the 14 months[2] of Blogs08) to mine the latent subtopics. The training data in time slice t_i is the top 2000 relevant English documents D_{t_i} returned by Okapi BM25 retrieval model, such that for a document $d \in D$, $pubDate(d) \in t_i$. The optimum number of subtopics is determined on-the-fly in the range from (lower bound) $\gamma = 5$ to (upper bound) $\delta = 20$. The subtopic models are trained using Mallet[3], an open source topic modeling tool, with its default parameters.

Preliminary analysis in the Blogs08 collection

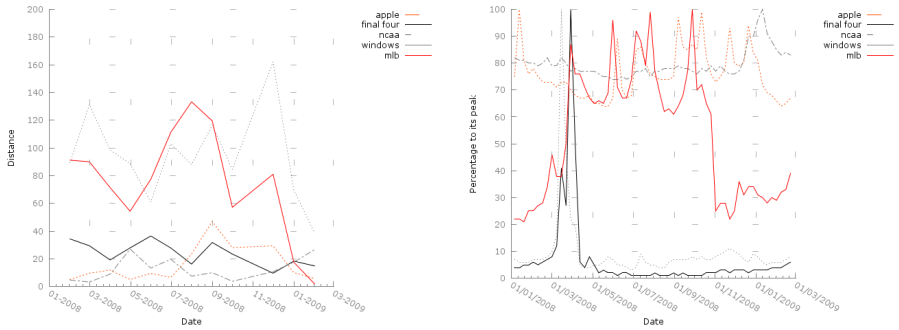
We adopt the metrics proposed by [12] to measure the similarity between two latent LDA topics in order to assess the dynamic or variance between subtopics mined from LDA from two consecutive time slices t and t+1. The idea is to measure the similarity between two latent topics modeled by LDA, ϕ_i^t at time slice t and ϕ_j^{t+1} at the consecutive time slice $t+1$.

Figure 2a demonstrates the dynamics of latent subtopics mined by LDA over time for the queries ncaa, windows, apple, mlb and final four. The dynamics of these queries are visualized by measuring the distance between two sets of subtopics mined from two consecutive time slices. We first observe the different level of subtopic dynamics of the four queries over time. The subtopic dynamics of query windows or apple are rather lower compared with ncaa or mlb. Kulkarni *et al.* [13] has initially analyzed the temporal correlations between the developments of query volume (or document content) and the query subtopics. In Figure 2b, we present the time series of these queries constituted using the query volumes retrieved from Google Trend[4] at the same time period with Blogs08. The query volumes are normalized by the peaked volume. We further observe that queries with low dynamic level tend to have stable high query volumes over time, i.e.,

[2] A finer-grained granularity can be achieved by mining from other source (e.g. query log), here we use monthly to acquire sufficient training data for our temporal queries

[3] http://mallet.cs.umass.edu/

[4] http://www.google.com/trends/



(a) **(b)**

Fig. 2: Temporal dynamics of LDA subtopics measured by (a) JS Divergence and (b) normalized query volumes from Google Trend

windows or apple while queries of high dynamic level (i.e., ncaa or mlb) tend to have sudden peaks in query volume. The query final four has a similar query volume development with ncaa, however, has a lower level. It is because final four only indicates a specific stage of the NCAA basketball tournament. The coverage of its aspects are then narrower than the query ncaa. A quantitative analysis of the correlation is left for future work.

Evaluation metric and result Perplexity is used to evaluate the quality of our trained LDA models (the ability of a model to generalize documents). The better the generalization performance of a model is, the lower is the perplexity score of its output (towards zero). We use holdout validation with 90% of the data for training and 10% for testing. We obtained the average perplexity of 0.00715 (variance is 0.000437).

4.3 Diversification Models

Due to the time gap between the AOL query log (March to May 2006) and Blogs08 collections (crawled from January 2008 to February 2009), we exclude the subtopics mined from AOL query log. The latent LDA subtopics mined from Blogs08 are the sole source of subtopics in this experiment. We take the top-10 words from the word probabilities of a LDA topic as an explicit representation of the subtopic. We evaluate the effectiveness of diversification models at diversifying the search results produced by Okapi BM25 retrieval model. Only English documents with content-length of more than 300 characters are accepted in the final top-100 results. For each query, we choose a studying time-point (as a simulated hitting time) based on the burst period of its query volumes derived from Google Trend (e.g., for US Open, the time-point is *June 2008* and *September 2008*).

Relevance assessments In this work, there is no existing gold standard dataset. Instead, we build our own gold standard on the Blogs08. From the top-100 documents for each of the (30) queries, we assess the subtopic-document relevance using human assessment. The relevance criteria is based on how relevant

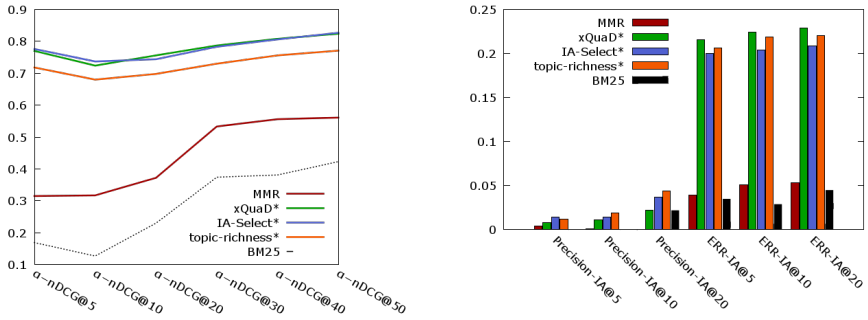


Fig. 3: Ranking results of baseline models, * models are with dynamic subtopic mining

is the document to the subtopic at the simulated hitting time. Each document is given a binary relevance judgment (by two experts), as follows the same setting from TREC Diversity Track 2009 and 2011. Given this orientation, a document is assessed based on the two dimensions, *relevance* and *time*. E.g., a document written about some happening that is content relevant to the subtopic but outdated is considered irrelevant. Notice that we asked the judges to assess with regards to different hitting times (simulated by monthly granularity)[5].

Evaluation metrics To evaluate the performance of our time-aware models, we use three different metrics (i.e., α-nDCG, Precision-IA and ERR-IA) that account for both the diversity and relevance of the results. In our evaluation, all metrics are computed following the standard practice in the TREC 2009 and TREC 2011 Web track [7,8]. In particular, α-nDCG is computed at $\alpha = 0.5$, in order to give equal weights to both relevance and diversity. We made a slight difference that in TREC 2009 Web track where they consider all query aspects equally important. We set the subtopics weight based on our dynamic subtopic measurement.

State-of-the-art model performance We measure the performance of the four state-of-the-art models: MMR, xQuaD, IA-Select and the topic richness model. The results are shown in Figures 3. For xQuaD, IA-Select and topic-richness, we use the mined temporal subtopics and their temporal weights as input (we skip their static methods (e.g., via Open Directory Project) since it is irrelevant in our case). We denote this change to the models with (*) symbol. We observe that measuring with α-nDCG@k, xQuaD*, IA-Select* and topic-richness model outperform MMR, while MMR shows certain increase over the baseline where there is no diversity re-ranking. We observe the same fashion when measuring with Precision-IA@k and ERR-IA@k. The results are expected since MMR does not account for subtopics when diversifying top-k result, it just tries to maximize the content gap between the top-k documents.

Time-aware parameter optimization The recency rate parameter λ is tuned to optimize the diversification models. We test λ in a wide range from 0.01 to 0.40. The parameter value with highest performance in terms of α-nDCG@k, ERR-IA@k and Precision-IA@k is chosen as best parameter value for the latter

[5] The judgment is available at: `www.l3s.de/~tunguyen/ecir2014_dataset.zip`

Table 4: α-nDCG results with $^{\triangle}$ ($p < 0.05$), $^{\triangle\triangle}$ ($p < 0.01$) indicate a significant improvement

	$\alpha - nDCG@5$	$\alpha - nDCG@10$	$\alpha - nDCG@20$	$\alpha - nDCG@30$	$\alpha - nDCG@40$	$\alpha - nDCG@50$
temp-xQuaD	**0.783**$^{\triangle}$	**0.737**$^{\triangle}$	**0.758**$^{\triangle}$	**0.805**$^{\triangle\triangle}$	**0.820**$^{\triangle}$	**0.847**$^{\triangle}$
xQuaD*	0.699	0.687	0.706	0.751	0.772	0.789
temp-IA-Select	**0.781**	**0.739**$^{\triangle\triangle}$	**0.755**$^{\triangle\triangle}$	**0.798**$^{\triangle\triangle}$	**0.822**$^{\triangle\triangle}$	**0.836**$^{\triangle}$
IA-Select*	0.738	0.698	0.718	0.760	0.790	0.807
temp-topic-richness	**0.697**	**0.662**	**0.686**$^{\triangle}$	**0.731**$^{\triangle}$	**0.753**$^{\triangle}$	**0.769**$^{\triangle}$
topic-richness*	0.654	0.638	0.660	0.702	0.727	0.741

experiments. We choose k to be 10 in this set of experiments, as 10 is the common cutoff level in relevant diversity tasks [7,8]. We obtained λ equals to 0.04 as the optimal value of the experiments.

Diversification performance In these experiments, we aim to evaluate our time-aware models to answer our stated research question whether taking time into account that favors recency can improve the performance of the state-of-the-art diversification models. Tables 4 and 5 represent the results of the state-of-the-art and our time-aware models for α-nDCG and the two metrics Precision-IA and ERR-IA at different cutoffs respectively. The results for α-nDCG show that temp-XQuaD significantly ($p < 0.05$) outperforms the state-of-the-art xQuaD all cut-offs (with $p < 0.01$ at k = 30). temp-xQuaD also achieves better results for Precision-IA and ERR-IA, however the results are not significant. One intuitive reason is that, different from α-nDCG that is influenced by the diversity of the top-k document result, Precision-IA and ERR-IA is more sensitive on document ranking, while we only test on the top-100 documents. The margin value can become significant when testing with top-1000 documents for the two metrics. Similar to temp-xQuaD, temp-IA-Select surpass IA-Select in overall, significantly outperforms the state-of-the-art IA-Select when measuring by α-nDCG at the cutoff k = 10, 20, 30, 40 and 50. temp-IA-Select also gives better yet not significant performance when measured by ERR-IA. However, temp-IA-Select does not surpass the original IA-Select for Precision-IA. The results of Precision-IA at cutoff k = 5, 10 and 20 show a slight decrease in performance of temp-IA-Select. We also report the results for temp-topic-richness and topic-richness in a similar fashion. Overall, our time-aware models exceed their originated state-of-the-art diversification models in most of the experimental settings. temp-xQuaD is the most consistent algorithm that outperforms xQuaD and gives better results among the six tested algorithms. On the other hand, even though surpassing the based model, temp-topic-richness gives a lower performance compared to the other two time-aware diversification models. However, the model is meant for taking subtopics from multiple sources, its performance could be enhanced if we account for other sources of subtopics (i.e., query log).

Table 5: Precision-IA and ERR-IA results with $^{\triangle}$ ($p < 0.05$) indicate a significant improvement

	P-IA@5	P-IA@10	P-IA@20	ERR-IA@5	ERR-IA@10	ERR-IA@20
temp-xQuaD	**0.010**	0.011	**0.029**	**0.214**	**0.218**	**0.232**$^{\triangle}$
xQuaD*	0.008	0.011	0.021	0.206	0.214	0.219
temp-IA-Select	0.010	0.010	0.027	**0.207**	**0.216**	**0.235**
IA-Select*	**0.013**	**0.013**	**0.034**	0.014	0.194	0.198
temp-topic-richness	0.010	0.011	0.030	**0.191**	**0.196**	**0.201**
topic-richness*	**0.011**	**0.017**	**0.040**	0.181	0.188	0.193

5 Related Work

Studying the temporal dynamics of subtopics has been addressed in some recent works [19,20]. Whiting *et al.* [19] considered event-driven topics as a prominent source of high temporal variable subtopics (search intent). They proposed an approach (in the absence query log) to present query intents by sections in the Wikipedia article. They further linked the temporal variance of intents (reflected by query volumes) with the change activity of the article sections. The proposed approach has certain limitations where the temporal dynamics and complexity in content structure of a Wikipedia article (where the subtopics are mined) is left un-tapped. Zou *et al.* [20], in another aspect, studied the effects of such subtopic temporal dynamics for the task of diversity evaluation. They conducted a small study on the Wikipedia disambiguation pages to analyze the changes in a subtopic popularity (the number of page views) over time. They concluded that such temporal dynamics impact the traditional diversity metrics for ambiguous queries, where the subtopic popularity is considered static over time. On the other hand, Berberich *et al.* [3] aimed to diversify search results over time, for those queries that are temporally ambiguous (i.e., the relevant time is un-known). Their proposed model, therefore, ignores the underlying intents of such queries and solely focuses on diversifying the relevant time periods of such queries. Styskin *et al.* [18] proposed a machine learning approach to identify recency-sensitive queries. Their large-scale experiments on real (recency-sensitive) queries show that promoting recent results (to the extent proportional to the query's recency level) to the result sets increases users' satisfaction.

6 Conclusions

In this paper, we studied the problem of diversifying search results for temporally ambiguous, multi-faceted queries. For such queries, the popularity and relevance of their corresponding subtopics are highly time-dependent, that is, the temporal dynamics of query subtopics can be observed. We determined dynamic subtopics by analyzing two data sources (i.e., query log and a document collection), which provides interesting insights for the identified temporal subtopics. Moreover, we proposed three time-aware diversifying methods that take into account the recency aspect of subtopics for re-ranking. The experimental results show that

leveraging temporal subtopics as well as recency can improve the diversification performance (diversity and relevance) and outperform the baselines significantly, for temporally ambiguous, multi-faceted queries.

References

1. R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *Proceedings of WSDM '09*, 2009.
2. R. Arun, V. Suresh, C. E. Veni Madhavan, and M. N. Narasimha Murthy. On finding the natural number of topics with latent dirichlet allocation: some observations. In *Proceedings of PAKDD '10*, 2010.
3. K. Berberich and S. Bedathur. Temporal diversification of search results. In *SIGIR 2013 Workshop on Time-aware Information Access (TAIA'2013)*, 2013.
4. D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.
5. J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of SIGIR '98*, 1998.
6. B. Carterette and P. Chandar. Probabilistic models of ranking novel documents for faceted topic retrieval. In *Proceedings of CIKM '09*, 2009.
7. C. L. A. Clarke, N. Craswell, and I. Soboroff. Overview of the TREC 2009 web track. In *TREC*, 2009.
8. C. L. A. Clarke, N. Craswell, I. Soboroff, and E. M. Voorhees. Overview of the TREC 2011 web track. In *TREC*, 2011.
9. N. Craswell and M. Szummer. Random walks on the click graph. In *Proceedings of SIGIR '07*, 2007.
10. Z. Dou, S. Hu, K. Chen, R. Song, and J.-R. Wen. Multi-dimensional search result diversification. In *Proceedings of WSDM '11*, 2011.
11. N. Kanhabua and K. Nørvåg. Improving temporal language models for determining time of non-timestamped documents. In *ECDL '08*, 2008.
12. D. Kim and A. Oh. Topic chains for understanding a news corpus. In *Proceedings of CICLing '11*, 2011.
13. A. Kulkarni, J. Teevan, K. M. Svore, and S. T. Dumais. Understanding temporal query dynamics. In *Proceedings of WSDM '11*, 2011.
14. F. Radlinski, M. Szummer, and N. Craswell. Metrics for assessing sets of subtopics. In *Proceedings of SIGIR '10*, 2010.
15. D. Rafiei, K. Bharat, and A. Shukla. Diversifying web search results. In *Proceedings of WWW '10*, 2010.
16. R. L. Santos, C. Macdonald, and I. Ounis. Exploiting query reformulations for web search result diversification. In *Proceedings of WWW '10*, 2010.
17. W. Song, Y. Zhang, H. Gao, T. Liu, and S. Li. HITSCIR system in NTCIR-9 subtopic mining task, 2011.
18. A. Styskin, F. Romanenko, F. Vorobyev, and P. Serdyukov. Recency ranking by diversification of result set. In *Proceedings of CIKM '11*, 2011.
19. S. Whiting, K. Zhou, J. Jose, and M. Lalmas. Temporal variance of intents in multi-faceted event-driven information needs. In *Proceedings of SIGIR '13*, 2013.
20. K. Zhou, S. Whiting, J. M. Jose, and M. Lalmas. The impact of temporal intent variability on diversity evaluation. In *Proceedings of ECIR '13*, 2013.