Combining Implicit and Explicit Topic Representations for Result Diversification

Jiyin He
J.He@cwi.nl

Vera Hollink
V.Hollink@cwi.nl

Arjen de Vries
arjen.de.vries@cwi.nl

Centrum Wiskunde en Informatica
Science Park 123, 1098XG
Amsterdam, the Netherlands

ABSTRACT

Result diversification deals with ambiguous or multi-faceted queries by providing documents that cover as many subtopics of a query as possible. Various approaches to subtopic modeling have been proposed. Subtopics have been extracted internally, e.g., from retrieved documents, and externally, e.g., from Web resources such as query logs. Internally modeled subtopics are often implicitly represented, e.g., as latent topics, while externally modeled subtopics are often explicitly represented, e.g., as reformulated queries.

We propose a framework that: i) combines both implicitly and explicitly represented subtopics; and ii) allows flexible combination of multiple external resources in a transparent and unified manner. Specifically, we use a random walk based approach to estimate the similarities of the explicit subtopics mined from a number of heterogeneous resources: click logs, anchor text, and web n-grams. We then use these similarities to regularize the latent topics extracted from the top-ranked documents, i.e., the internal (implicit) subtopics. Empirical results show that regularization with explicit subtopics extracted from the right resource leads to improved diversification results, indicating that the proposed regularization with (explicit) external resources forms better (implicit) topic models. Click logs and anchor text are shown to be more effective resources than web n-grams under current experimental settings. Combining resources does not always lead to better results, but achieves a robust performance. This robustness is important for two reasons: it cannot be predicted which resources will be most effective for a given query, and it is not yet known how to reliably determine the optimal model parameters for building implicit topic models.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval

General Terms

Algorithms, Experimentation

Keywords

Multi-source, Subtopics, Result diversification, Random walk

SIGIR'12, August 12–16, 2012, Portland, Oregon, USA.
Copyright 2012 ACM 978-1-4503-1472-5/12/08 ...$5.00.

1. INTRODUCTION

Queries in Web search are often short and underspecified [2]. For example, the query "python" may refer to a snake as well as a programming language, while the programming language "python" covers a wide range of subtopics, such as tutorials, documentations, and downloads. Without knowing the users' actual intent, result diversification deals with such queries by providing documents that cover as many subtopics of a query as possible, so that the average user's "unhappiness" is minimized [43].

Result diversification has been intensively studied and many diversification algorithms have been proposed [1, 7, 8, 21, 34, 35, 43]. This paper focuses on the discovery of subtopics, also referred to as query intent, facets, and sub-queries in the literature. Subtopics have been extracted internally, e.g., from documents retrieved in response to the query [7, 8, 26, 29], and externally, e.g., from Web resources such as query logs [1, 20, 21, 33, 35]. Internally modelled subtopics are often implicitly represented, e.g., as latent topics or document clusters, while externally modelled subtopics are often explicitly represented, e.g., as reformulated queries[1]. In particular, there has been much work that exploits explicit subtopics using various types of Web resources. Some take suggested queries from commercial search engines as subtopics of a query [35]; others mine the subtopics from resources such as taxonomies [1, 22], query logs[21, 33], anchor text and Web ngrams [20].

While both implicit and explicit subtopics extracted from various Web resources have shown their effectiveness in helping retrieving diverse search results, both topic representations have their limitations. Often, the process of extracting subtopics for a query in a Web resource is focused on finding the subtopics that are relevant to the query, while the relation among the extracted subtopics are not well preserved. For instance, all subtopics discovered in Web resources, e.g., all reformulations of a query, are treated as independent subtopics, regardless the fact that some may be synonymous [21, 35]. In some studies additional steps such as clustering were taken in order to avoid redundancy and find better defined topics [20, 33]. On the other hand, implicitly represented subtopics such as topic models or clusters constructed from document content convey much information about the relations or the structure of the topics present in the documents. Yet, in this case the infor-

[1] Notice that in the context of result diversification, "implicit" sometimes refers to diversification methods such as MMR where no actual topics are modeled and only document similarities are used. In this paper, from a topic modeling perspective, we refer to topic models/clusters as implicit *topic representations*, since no "labels" that indicate the content of the topics are assigned, and refer to subtopics represented with external entities such as query suggestions, anchor text, as explicit topic representations, as such entities present the content of the topic explicitly [25].

mation that can be used is limited to the content of the document. Therefore we are interested in the following question:

- Can we make use of the information from both implicitly and explicitly represented subtopics?

Various resources were shown to provide useful information in subtopic mining, sometimes complementary to each other. Click logs have for example been shown an effective resource for subtopic mining in a number of studies [21, 33], but their availability are mainly limited to commercial web engines. Anchor text has been suggested a good proxy for query information, but it is not clear if one should be preferred over the other. Web ngrams are freely available, but may not provide as high quality subtopics for a query as click logs, partially because of the noise introduced by phrases not relevant to the topic of interest. Given a variety of resources, we are interested in a second question:

- Can we combine multiple external sources to help subtopic modeling?

We seek a principled way to model the subtopics of a query using multiple sources, including implicitly represented topics extracted from document content, as well as explicitly represented subtopics mined from various Web resources. While combining subtopics from multiple sources can be useful, it is non-trivial. Different resources provide evidence that a subtopic is relevant to a query from different perspectives, which leads to different types of measurements to quantify the similarity between subtopics. We propose a graph based approach to combine explicit subtopics from Web resources. More specifically, we construct local graphs over the subtopics extracted from each single resource. The local graphs are interconnected through subtopics that appear in multiple resources. A measure of subtopic similarity within each local graph is converted into random walk transition probabilities, to obtain a probabilistic framework. Subtopics from different sources are combined in a random walk over the resulting graph, that integrates the various external resources. The result of the random walk defines a probability distribution that encodes the similarity between all subtopics from all resources included in the graph. Given the similarities in the explicit subtopic graphs, we exploit these similarities for regularization during the construction of implicitly defined subtopics from the search result list. Using regularized topic models, the similarities among the explicit subtopics serve as additional constraints, that need to be satisfied when constructing implicit subtopics using document content.

Our contribution is two-fold. First, we propose a framework that: i) combines both implicitly and explicitly represented subtopics in a principled way; and ii) allows flexible combination of multiple external resources, in a transparent and unified manner. With respect to the proposed framework, an in-depth investigation on its effectiveness as well as limitations are provided. Second, using this framework, we compare the usefulness of various resources (and their combinations) for identifying diversification subtopics. Our findings provide additional evidence for findings in earlier studies on similar themes, and lay a foundation for future work on selecting useful resources for topic modeling.

The rest of the paper is organized as follows. In Section 2, we discuss the related work on subtopic modeling in result diversification, as well as work related to the techniques employed in this paper. In Section 3, we introduce our proposed framework, referred to as Multi-source subtopics (MSS), to combine explicit subtopics from multiple sources with the implicit ones derived from the initial search result list. In Section 4, we discuss the different types of resources considered for subtopic modeling. We then describe the experiments designed to empirically study the properties of the proposed framework in Section 5, followed by a discussion on the experimental results and their implications in Section 6. Section 7 concludes the paper.

2. RELATED WORK

Search result diversification has a long history [5, 7, 10, 23, 37, 43] and a range of diversification approaches have been proposed previously. Maximum Marginal Relevance (MMR) [7], an early representative diversification method, aims to balance the relevance and diversity of a ranked list. A probabilistic version of MMR has been proposed by Zhai et al. [43], as part of their risk minimization framework. Zhu et al. [44] proposed a ranking method based on random walks in an absorbing Markov chain. By turning ranked items into absorbing states, the method prevents redundant items from receiving a high rank. Yue and Joachims [42] studied a learning algorithm based on a structural Support Vector Machine (SVM) that identifies diverse subsets in a set of documents. Carterette and Chandar [8] proposed a probabilistic model for faceted retrieval, where the facets of a query are modeled with Latent Dirichlet Allocation (LDA) [4], one of the state-of-the-art topic modeling approaches. In most of these studies, the focus has been on the development of ranking algorithms to produce ranked lists that convey both relevant and diverse information about a query. Only limited effort has been made in mining high quality subtopics for a query, and subtopics were usually modeled based on document content.

Recently however, in the context of Web retrieval, a renewed interest in result diversification has emerged. Representative approaches include Intent Aware select (IA-select) [1] and explicit query aspect diversification (xQuAD) [35]. Common to these approaches is the importance of modeling both the subtopics of a query *and* the relevance of a document with respect to the subtopics of a query. Both works exploited Web resources for subtopic mining. In [1], categories in the Open Directory Project (ODP) taxonomy were used as candidate subtopics, and in [35], query suggestions from Web search engines were used as related subtopics of a query. Recently, Dou et al. [21] has first proposed to combine subtopics mined from multiple sources, including query logs, anchor text, document clusters, and web sites of search results. A common assumption among studies where Web resources are used for subtopic extraction is that the extracted subtopics would be independent. However, these assumptions are most likely not true. Anchor text has for instance been shown to be an effective substitute for query logs in generating query reformulations [19], which indicates that these two resources should be correlated. Moreover, within a single resource such as query logs, various reformulations of a query may be synonyms that refer to the same concept. Our approach does not require to assume the independence among mined subtopics. Conversely, we focus on modeling specifically the similarity or relatedness among the subtopics. In our experiments, we employ IA-select as the primary diversification method (see Section 5.3), to indirectly evaluate the effectiveness of using our approach to model subtopics from multiple sources.

A closely related line of research focuses on mining subtopics of a query from various Web resources. Radlinski et al. [33] infer query aspects from query logs, using clicks and session information to model the relations between queries. Dang et al. [20] infer query intent from anchor text and Web ngrams. In both studies, clustering has been applied to the extracted aspects or intents, so that topically redundant or similar entities (queries, anchor text, or Web ngrams) are grouped together. In this paper, we use the same types of Web resources for subtopic mining, and compare their effectiveness in the context of result diversification; although in terms of a query log resource, our data set can only be seen as a very small subset of the log used in [33]. Instead of applying clustering, we incorporate the obtained similarities between the extracted subtopics into the topic models constructed using document content. This way we incorporate the similarity between extracted subtopics into the similarity between documents.

To compute the similarity between subtopics extracted from different Web resource, we construct graphs over these subtopics and compute their associations using Markov random walks. Similar approaches have been studied in a wide range of topic in the context of query log analysis. For example, Craswell and Szummer [16] studied the usage of Markov random walks over the bipartite graph constructed from click logs for document ranking. Fuxman et al. [22] developed a random walk algorithm for keyword generation from query click graphs. This approach was also adopted in [1] to extract subtopics related to a query from the ODP taxonomy. Ma et al. [30] proposed a query suggestion approach based on Markov random walk and hitting time on the query-URL bipartite graph, aimed at suggesting both semantically relevant and diverse queries to Web users. A recent work that is closely related to ours is the Multi-view random walk (MVR) approach proposed by Cui et al. [18], which aims to combine knowledge from different views of an object in a random walk process. We discuss the differences between their work and ours in detail in Section 3.

To balance graph-based and textual similarity between documents, we use regularized topic models [6, 24]. A similar approach was taken in Guo et al. [24], where queries were enriched with document snippets to improve a query similarity model; we however enrich the document content with queries to construct a document similarity model.

3. MULTI-SOURCE SUBTOPICS

3.1 Modeling relation among explicitly represented topics from Web resources

We now detail how we use a Markov random walk based approach to compute the relations between the explicit subtopics of a query in different types of resources. Let $R = \{r_i\}_{i=1}^N$ be a resource that contains N subtopics related to a query q. Consider for instance a query log that contains N reformulations of q. For a set of resources $\mathcal{R} = \{R^g\}$, we construct a network consisting of subgraphs on multiple parallel 'planes', as shown in Figure 1. First, for each R^g, we construct a weighted graph $G^g = \{E^g; R^g\}$, where the nodes R^g corresponding to the subtopics in resource g lie on a single plane, and the edges E^g are weighted by the similarity $w(r_i, r_j)$ between the two nodes. For now, we abstract from the specific way the similarity is defined within a plane, aiming for a method that can be applied generically; details of the graphs created for specific resources are deferred to Section 4. In the second step, we interconnect nodes from different planes G^a and G^b with an extra set of edges, where each edge connects the nodes $r_i \in G^a, r_j \in G^b$ that we consider equivalent. Throughout this paper, we consider r_i and r_j equivalent whenever they have exact matching text representations. While it would be interesting to include non-exact matches, we leave this for future investigation.

For each resource considered, the similarity scores computed may encode different semantics, such that they may not be compared directly across different planes. This is the primary reason why we are not in favor of constructing a single graph that mixes up subtopics from different resources. To resolve this problem, we first transform the similarity scores into 1-step random walk transition probabilities. The only assumption behind the resulting transition probability is that the chance to travel from one subtopic to another in a random walk step should depend on the similarity between these subtopics, i.e., the more similar two subtopics are, the more likely the transition from one subtopic to the other can happen.

By applying a t-step random walk over the graphs, we expect to find new associations, in particular between subtopics that were not directly connected in a single graph. Further, the random walk sums up the transition probability of all paths of length t between two nodes. The transition probability will be high if there exist many paths from one node to the other. The higher the transition probability the more similar or closely related the two nodes are. During the random walk, the underlying clustering structure of the nodes emerges [36]. If such a clustering structure exists among the nodes, with longer walks, the probability between nodes within a cluster converges to a value higher than that with any node outside the cluster [16]. The probability distribution that arises can therefore be interpreted as a similarity measure that captures the relations between the subtopics in our graphs.

We use $p_1^\theta(r_j|r_i)$ to denote the 1-step transition probability from r_i to r_j within a single plane, which is computed as

$$p_1^\theta(r_j|r_i) = w(i,j) / \sum_j w(i,j). \tag{1}$$

We refer to the nodes connected to nodes in other planes as "teleport points" and the transition from one plane to another as a "between-plane teleport". We use $p_1^\beta(r_j|r_i)$ to denote the 1-step between plane transition probability from r_i to r_j. While it is possible to assign a different value for each pair of nodes, e.g., based on additional evidence or the performance on training material, in this paper we only study the simplified situation assuming each plane has a fixed p_1^β. p_1^β can be viewed as the probability that a plane is chosen as the destination of a teleporting. For simplicity, we use β_g to denote this probability for plane g. In summary, we have

$$p_1^\beta(r_j|r_i) = \begin{cases} 0 & \text{if } r_i \text{ is not a teleporting point,} \\ \beta_g & \text{otherwise, where } r_j \in G^g. \end{cases} \tag{2}$$

Piecing together the within-plane and between-plane transition probability, while conducting the random walk, the 1-step transition probability from r_i to r_j is computed as follows.

$$p_1(r_j|r_i) = \begin{cases} p_1^\theta(r_j|r_i) & \text{if } r_i, r_j \in G^g, \\ p_1^\beta(r_t|r_i) p_1^\theta(r_j|r_t) & \text{otherwise,} \end{cases} \tag{3}$$

where r_t is the teleport point in the plane of r_j connected with r_i.

A t-step walk from r_i to r_j is defined recursively as

$$p_t(r_j|r_i) = \sum_k p_1(r_j|r_k) p_{t-1}(r_k|r_i), \tag{4}$$

where r_k's are the intermediate nodes that directly connect to r_j.

At first sight, our proposed representation may look similar to the multi-view random walk (MVR) model [18]. In MVR however, each plane represents a different view of the exact same set of objects; i.e., each object within the network has a corresponding node at each plane. In our setup, this correspondence is not required, and therefore a phrase occuring in, for example, an anchor text collection does not necessarily have to occur in a query log as well. Not requiring such a correspondence allows for a greater diversity of resources used. Also, MVR defines a 1-step cross-view transition only between corresponding nodes. In principle, although we have not yet studied this alternative setting, our representation allows to consider also a between-plane 1-step transition from a node in one plane to multiple nodes in the other plane, i.e., setting the $p_1^\beta(j|i)$ with non-zero values to each pair of nodes i and j.

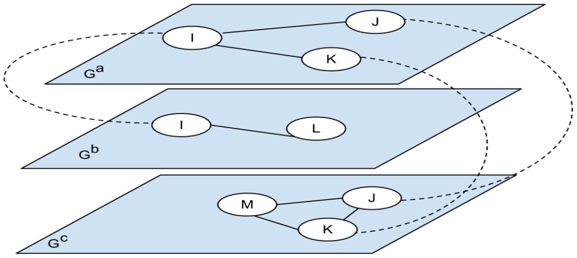


Figure 1: An illustration of the mutli-source subtopic network. Solid lines indicate within plane transitions, and dashed lines indicate the between plane teleporting. Each of the nodes I, K, J, L, M, are possible intents for the same query. Each plane represents intents generated from a single data source.

3.2 Combining implicit and explicit subtopics

Topic models, such as the probabilistic latent semantic analysis (pLSA) [28], are commonly used to model underlying topics associated with a set of documents. The extracted topics are implicitly represented by a set of word distributions. For a query q, we apply pLSA on the set of retrieved documents $D = \{d_i\}_{i=1}^M$ to obtain the implicit subtopics associated with q.

Now that we have described our approach to model the relations between subtopics extracted from multiple resources, the next question is: how can we combine the relations between the explicit subtopics with the implicit subtopics? Notice that for diversification, the ultimate goal is to generate a ranked list of *documents* that are both relevant to the query and contain diverse content. Therefore it is necessary to associate the mined subtopics to the documents to be ranked, and convert the relations between subtopics to the relations between documents.

We first translate similarities between explicit subtopics to the similarities between documents as follows.

$$p(d_i|d_j) = \sum_{k,l} p(d_i|r_k)p_t(r_k|r_l)p(r_l|d_j) \tag{5}$$

To obtain $p(r|d)$, we use r as a query and compute the query likelihood score as defined in a language modeling retrieval model [32]. Then following the Bayes' rule, $p(d|r)$ is computed as

$$p(d_i|r_k) = p(r_k|d_i)p(d_i)/p(r_k), \tag{6}$$

where $p(r_k)$ is computed by marginalization as $\sum_{d_i \in D} p(r_k|d_i)$, and $p(d_i)$ as $1/|D|$, assuming a uniform distribution of $d \in D$.

Given these preliminaries, we now combine the similarities between documents, obtained from the explicit subtopics, with the implicit topics. Hereto, we apply Laplacian pLSA [6] (also referred to as regularized topic models [24]), using the document similarities given by Eq. 5 to regularize the implicit topic model. Laplacian pLSA employs a generalized version of EM to maximize the regularized log-likelihood of the topic model, $\mathfrak{L}$:

$$\mathfrak{L} = \mathcal{L} - \gamma\frac{1}{2}\sum_{k}\sum_{i,j}(P(z_k|d_i) - P(z_k|d_j))^2 p(d_i|d_j) \tag{7}$$

Here, $\mathcal{L}$ is the log-likelihood of the implicit topic model as maximized by pLSA. $P(z_k|d_i)$ is the probability of topic z_k given document d_i. γ is a parameter that controls the amount of regularization from external resources.

By maximizing the regularized log-likelihood, Laplacian pLSA (softly) assigns documents to the same cluster if they 1) share many terms *and* 2) belong to the same explicit subtopics. γ allows us to balance these two requirements and combine both implicit and explicit representations of query subtopics in a unified and principled manner. Notice that when no explicit subtopics can be found for a query, the regularized pLSA is reduced to the normal pLSA.

4. RESOURCES

We now describe the implementation of each plane of the network. In this paper, we consider three resources: click logs, anchor text and Web ngrams (summarized in Table 1).

Graph	Nodes	Edge weights
G^C	search queries	coclicked documents
G^A	anchor texts	coocurrence in text passages
G^N	web ngrams	coocurrence in text passages

Table 1: Used resources

4.1 Click logs

Clickthrough data has proven to provide useful information for identifying search intents, or, related topics of queries [24, 33]. In particular, if two queries share many co-clicked documents, it is likely that they convey similar topical information [3, 39].

We construct a graph G^C from click logs, using logged queries as nodes, following a rather standard approach of using the total number of co-clicked documents to set the edge weights for random walk (see e.g. [16]). Since our goal is to find subtopics related to a query, unlike in many other studies where a global graph is constructed over all logged queries, we construct local graphs for a given query. The advantage of using local graphs is two-fold: i) it is computationally efficient, and ii) queries that are not relevant to the original query are effectively pruned. In order to collect related subtopics for a query q, we find the queries that share co-clicked documents with q, and for each of these queries further find its co-clicked queries. We then filter out queries that has a 1-step (or 2-step in the latter case) transition probability less than 0.01 starting from q in order to prevent popular queries such as "yahoo" from connecting to many weakly related queries.

4.2 Anchor texts

Anchor text has proven to be another effective feature for various search tasks, e.g. [19]. We use the anchor texts from the ClueWeb09 collection [27] to construct an anchor text graph G^A, using the method described by Dang et al. [20]. As preprocessing, we remove all anchors that are connected to only one URL. For a given query, we collect the N (N = 100 in our experiments) most frequently occurring anchors that contain all of the query terms. These anchors become the nodes of the graph. The edge weights are computed according to the method that proved most effective in the study of Dang et al. [20]: on the basis of the number of times the anchor phrases co-occur in text passages in the collection.

4.3 Web n-grams

Following Dang et al. [20], we collect Web n-grams using the Microsoft Bing N-gram [38] service to build an n-gram graph G^N. For each query q, we retrieve the M (M = 1000) terms t with the highest probability of seeing t after query q in the body text of the web pages indexed by Bing, $p_{ngram}(t|q)$. As the N-gram service provides n-grams up to length 5, this graph can be constructed for queries consisting of 4 or less terms. Dang et al. [20] use the N terms with the highest $p_{ngram}(t|q)$. However, these often include common terms that are not specific to the queries. Therefore, we

Sample subtopics	Top 3 related subtopics					
G^C	subtopic	score	subtopic	score	subtopic	score
anti spy	windows defender	.2261	microsoft antispyware	.1208	defender	.1122
microsoft spyware	windows defender	.2262	microsoft antispyware	.1208	defender	.1121
antispyware	windows defender	.2265	microsoft antispyware	.1207	defender	.1121
microsoft beta	windows defender	.2260	microsoft antispyware	.1209	defender	.1120
windows defender	microsoft antispyware	.1218	defender	.1141	antispyware	.0995
G^{CA}	subtopic	score	subtopic	score	subtopic	score
space defender 1 0	star defender 4	.1266	star defender 3	.1266	star defender 2	.1266
defender industries	defender industries inc	.2055	defender	.1197	windows defender	.0462
microsoft beta	windows defender	.1062	microsoft defender	.0555	microsoft s windows defender	.0538
a public defender	public defender	.1160	public defender's office	.1040	office of the public defender	.1040
tri state defender	chicago defender	.1035	the chicago defender	.1035	national legal aid defender association	.0352

Table 2: A random sample of 5 subtopics related to the query "defender" from G^C and G^{CA} and the top 3 subtopics related to each of the sample subtopics. The scores are the result of a 5-step random walk on the corresponding graphs.

compute for each retrieved term the lift [41]:

$$lift(t, q) = p_{ngram}(t|q)/p_{ngram}(t) \qquad (8)$$

where $p_{ngram}(t)$ is the a priori probability of seeing term t. The N (N = 100) n-grams (= query + term) with the highest lift are included as nodes in the graph. Following Dang et al. [20], the weights of the edges are computed in the same way as in G^A.

4.4 An example of Multi-source subtopics

Having introduced the MSS framework and the target Web resources, now let us see an example that illustrates the result of the random walks on the subtopic graphs.

Table 2 shows an example result of a 5-step random walk on two different graphs, G^C and G^{CA}, i.e., the graphs constructed using click logs, and the combination of click logs and anchor text, respectively. We take the query "defender" and randomly sample 5 subtopics related to this query from each graph. For each sampled subtopic, we list the 3 most similar subtopics (from left to right), where the similarity is measured by the transition probabilities starting from the sampled subtopics as a result of the random walk. In this example, G^C contains 21 subtopics, and G^{CA} contains 118 subtopics. We see that all the subtopics sampled from G^C are closely related to the security software "windows defender", which indicates that this is a dominant topic in G^C. The sample drawn from G^{CA}, on the other hand, covers diverse topics: computer games (space defender, star defender, etc.,), defender industries, windows defender, public defender, and subtopics about the newspapers "tri-state defender" and "chicago defender". On top of that, if we take a close look at the resulting similarity scores, we see that these scores effectively reflect the semantic relatedness between the subtopics. For example, "defender industries" is closely related to "defender industries inc", as both can be interpreted as the marine and boat supply company "Defender Industries Inc."; it is loosely related to "defender", as it is the original query, which is vague and can be interpreted as anything; it is not very likely to be related to "windows defender", and correspondingly, a very weak relation is indicated by the similarity score.

5. EXPERIMENTS

5.1 Research questions and experimental setup

In this section, we empirically investigate the properties of the proposed framework and its impact on result diversification.

Our framework combines different types of resources at two levels. First, it combines the implicitly and explicitly represented subtopics by constructing regularized topic models. It is therefore natural to investigate the following research question.

RQ1 Does regularization with subtopics extracted from external resources help to form better topic models?

Although the question asks whether "better topic models" are formed, it is difficult to access the quality of topic models directly, due to the subjective nature of topics. Instead, by applying the resulting topic models to diversification, we indirectly assess their quality by examining the diversification performance. It is reasonable to assume that better topic models lead to better diversification results.

Further, the framework combines explicit subtopics from multiple external resources, which leads to the following two questions.

RQ2 How do various subtopics from external resources and their combinations compare in terms of diversification performance?

RQ3 Are combinations of subtopics from different external resources more effective in terms of diversification performance than that of single resources?

Particularly, previous studies suggest that anchor text is an effective replacement for query logs (which are often not publicly available), e.g., for query reformulation [19] and inferring query subtopics [20]. Comparison of the two resources within our framework can provide supplementary evidence and validate these findings.

Note that our framework outputs regularized topic models of a query, i.e., an implicit topic representation. Like any topic model based approach, LapPLSA (Laplacian pLSA) depends on a prefixed parameter, the number of topics K. There is no easy solution to find the optimal K without prior knowledge or sufficient training data. In our case, neither is available. However, in reality, "neither is available" is often the practice. Hence, it is useful to investigate the robustness of the framework when K is not optimal.

RQ4 How sensitive is the performance of diversification based on LapPLSA to the choice of K?

5.2 Data

We use the Clueweb09 setB dataset and the topic sets released at TREC2009-2011 Web Track diversity task as our test collection [12–14]. Since the topic sets are designed to be different, e.g., the 2011 topic set is expected to contain "tougher" queries [14] than the others, we experiment with topics from each set separately.

The following data are used as the Web resources that provide query subtopic information. We use the MSN query log [17] to construct G^C. Notice that this log is fairly limited: it contains 15 million queries from US users, sampled over 1 month. Following [20],

Graph	Coverage		
	1-50	51-100	101-150
G^C	39	37	21
G^A	48	47	25
G^N	48	45	34
G^{CA}	48	48	31
G^{CN}	50	48	39
G^{AN}	50	48	39
G^{CAN}	50	48	39

Table 3: Coverage of Web resources over the TREC topic sets

we use the Microsoft Ngram service to construct G^N, and anchor texts extracted from Clueweb09 dataset to construct G^A[27].

Table 3 shows the coverage of the Web resources on the TREC topics. A TREC topic is "covered" by a resource if the corresponding graph of the query is not empty. We see that in terms of individual resources, click logs has the lowest coverage on all three topic sets. In terms of topic sets, the set from TREC 2011 is obviously poorly covered. Combining multiple resources increases the coverage over all three sets, but still, not all topics are covered.

5.3 Diversification method

After pilot experiments with a number of state-of-the-art diversification methods, including IA-select [1], xQuAD[35] and MMR [7][2], we decide to focus on IA-select. Results between IA-select and xQuAD are comparable, while MMR has a lower performance. The advantage of IA-select is that it has no extra parameters to be tuned: both xQuAD and MMR have an extra parameter that linearly combines relevance and diversity. IA-select provides a more transparent way to analyze the behavior of our framework.

Given a set of candidate documents and a set of subtopics related to a query q, IA-select [1] selects a document d to be included in the ranked list base on: the relevance of d to q, and the probability that d covers subtopics that all previously selected documents failed to do so. The key elements used in the algorithm can be reduced to the following quantities: i) $V(d|q,z)$, the probability that d is relevant to q when the intended subtopic is z; and ii) $p(z|q)$, the probability that subtopic z is related to q.

The first quantity is determined by the retrieval score of d given q, weighted by the likelihood that d covers subtopic z, i.e., $p(d|z)$, which can be derived from the resulting topic models.

We compute $p(z|q)$ as follows.

$$p(z|q) = \sum_d p(z|d)p(d|q), \tag{9}$$

where $p(z|d)$ is given by the topic models, and

$$p(d|q) = \sum_r p(d|r)p(r|q). \tag{10}$$

where r are the nodes in the graphs; $p(d|r)$ is calculated using Eq. 6 and $p(r|q)$ is given by the result of random walks. Since our graphs were constructed locally using the original query q as the starting point, all graphs will at least contain the original query as a node. If the external resources do not contain any subtopics related to q, Eq. 10 reduces to the relevance score of d to the original query q.

We use IA-select to re-rank a pre-retrieved set of documents. To create the baseline ranked list, we use the Indri toolkit that implements the retrieval model based on the language modeling framework [31]. We use the spam list released at TREC Web Track [15] for spam filtering, setting the spam threshold to 70%. The top ranked 100 documents are used to construct topic models.

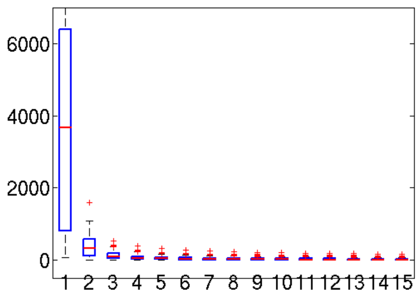


Figure 2: Change of probability distributions during random walks. X-axis shows the number of steps; Y-axis shows the KL-divergence between the states before and after a 1-step walk.

5.4 Parameter settings

Our proposed framework comes with a number of parameters. For random walks, two parameters are involved, namely β, the between plane transition probabilities, and t, the number of steps for the random walks[3]. For LapPLSA, we have: the regularization parameter γ, and the number of topics K.

As discussed in Section 3.1, β is practically the probability that a plane is chosen as the destination of a teleporting. As we do not have evidence or prior knowledge about whether one plane should be preferred over others, we simply assume a uniform distribution over all planes: $\beta_g = 1/|g|$, where $|g|$ is the total number of planes.

Since all our graphs are fairly small due to the fact that they are constructed locally based on subtopics that are closely related to a given query, random walks on these graphs converge very quickly. With a few preliminary experiments, we find that within 5 steps, the changes of the probability distributions over most graphs drop to a level that is almost negligible. Therefore we set t to 5 steps. A typical example is shown in Figure 2, where we use KL-divergence to measure the change of probability distributions over a graph. The graph consists of subtopics from click logs, anchor text and Web ngrams. The box plot is constructed over 150 TREC topics.

Regularization parameter γ can take the value of non-negative real numbers; Researchers have usually set its value empirically [6, 24]. Preliminary experiments with our method suggest that the end-to-end diversification results are relatively stable when varying γ within a range around 10, although fine tuning γ can indeed lead to improved results for a specific experiment. We decide to set γ to a fixed value that generates reasonable diversification results, using $\gamma = 10$ in all our experiments. Finally, note that $\gamma = 0$ makes LapPLSA equivalent to pLSA without regularization.

While results are relatively stable with respect to γ, we find that the performance of diversification with topic models is rather sensitive to the parameter K. In Section 6, we will discuss the impact of K on the diversification results using our framework.

5.5 Evaluation metrics

We evaluate the diversification result in terms of α-nDCG@20 [11] and ERR-IA@20 [9], which are primary evaluation metrics at TREC Web Track for the diversity task [14]. α is set to 0.5 for α-nDCG@20. For significance testing, we use the Wilcoxon sign rank test. When

[2] Unlike IA-select and xQuAD, MMR does not actually require subtopics being modeled. Here, when computing the similarities between documents, documents are represented by topic distributions in stead of term distributions.

[3] The implementation of the MSS random walks can be downloaded at `http://code.google.com/p/mss-rw/`

reporting results, we use $^{\triangle}$($^{\blacktriangle}$) to indicate a significant difference with p-value<.05 (.01).

6. RESULTS AND DISCUSSION

We now discuss the experimental results and their indications to the answers to our research questions. In the following discussion, we will first discuss the behavior of LapPLSA regularized by subtopics from different types of external resources in terms of the general trends shown by the results. We then zoom in to a few specific settings of K and examine the effectiveness of the proposed framework, e.g., highlighting the significance of the observed differences between approaches. Although the number of topics K is an important parameter that has an impact on the diversification result, we do not attempt to optimize it. We do investigate the sensitivity of our results to the estimation of K, when discussing RQ4.

6.1 Single resources

We start with the performance of LapPLSA using single resources. Figure 3 shows the result of IA-select using topic models constructed with the following methods: pLSA without regularization and LapPLSA regularized by similarity matrices generated using click logs, anchor text, and Web ngrams, i.e., LapPLSA_C, LapPLSA_A, and LapPLSA_N, respectively. “Baseline” refers to the run without diversification.

First, we see that both pLSA and LapPLSA (with different resources) can outperform the baseline. Compared to pLSA, LapPLSA shows more robust performance: diversification with pLSA can underperform the baseline given an improperly set K, while diversification with LapPLSA regularized by the subtopics from an external resource in general outperforms the baseline irrespective of the choice of K. The only exception is the case where $K = 2$, which is presumably not a sensible choice for K.

Second, judging from Figure 3, the effectiveness of each resource differs on different topic sets. It is noticeable that on topic set 1-50, click logs remarkably outperform the other two resources across all settings of K. A possible explanation is that this topic set is derived from query logs of commercial search engines [12], and therefore the click logs have a relatively high coverage and turn out to be an effective resource for these topics. On topic set 51-100, click logs and anchor text show comparable performance, while Web ngrams are occasionally effective (given a specific settings of K). On topic set 101-150, anchor text generally outperforms click logs. This may due to the fact that the click logs have a very low ($< 50\%$) coverage on this topic set, and that the topic set is rather recent (created in 2011) while the click logs were created in 2006, which may lead to further sparseness: e.g., on average, G^A has 17.1 nodes per query, while G^C only has 7.6 nodes per query on this topic set. In general, click logs and anchor text seem to be more valuable resources for regularization compared to Web ngrams, across different settings of K. Notice that the Web ngrams are primarily derived from document content, so perhaps their lower effectiveness can be explained by lower influence on pLSA, which also uses document content. To some extent, we can consider the Web ngrams more similar to the document content than click logs and anchor text.

As expected, the diversification results of IA-select based on both pLSA and on LapPLSA are sensitive to the change of the parameter K. In particular, there is no clear correlation between the number of clusters and the end-to-end diversification performance, which further suggests the difficulty of finding an optimal K (that would fit for a set of queries). For more detailed analysis on the parameter K, see Section 6.4.

6.2 Combining multiple resources

Figure 4(a)-4(c) show the result of combining subtopics from the two relatively more effective resources, namely click logs(G^C) and anchor text (G^A). While we only show the combination of these two resources, a similar trend can be observed in the other combinations. We see that combining resources does not always lead to improved diversification results over that of the single resources. Such improvement only happens in a few cases where K is likely to be in an optimal setting for the combined resource G^{CA}, i.e., $K = 10$ in Figure 4(b) and $K = 6$ in Figure 4(c). In general, the combination of different resources leads to an “averaged" performance compared to the individual resources that are combined.

This observation is reasonable. Intuitively, combining resources on the one hand increases the coverage of subtopics, e.g., as shown in Table 3, but on the other hand may introduce noise, e.g., if one of the resources contains low quality subtopics. Recall that when constructing the local graphs, their nodes (i.e., subtopics) were selected based on their relation to the original query. In this stage, the relevance of the nodes with respect to the original query was taken into account. However, when combining resources, i.e., during the random walk stage, the goal was to measure the similarities among nodes across resources and no further pruning was conducted. That is, all entries in different graphs were assumed to be equally good. Given our observations on the combined result, a natural step for future work would prune further to prevent low quality resources from deteriorating high quality resources.

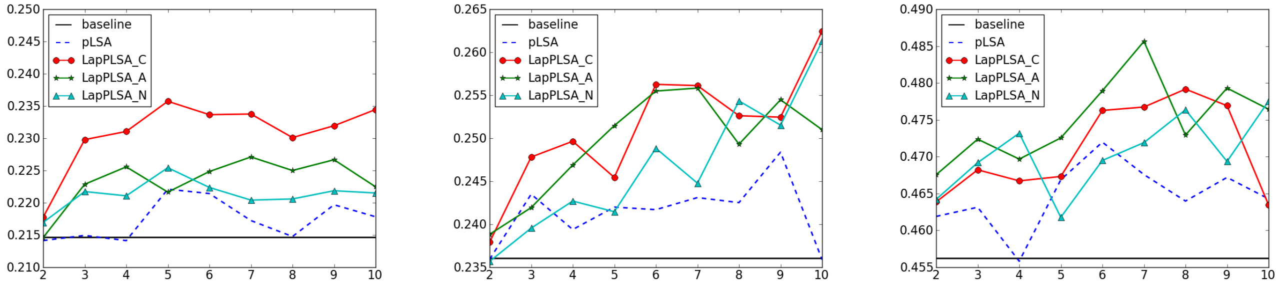
6.3 Zoom in on specific settings of K

Now, let us zoom in to the performance of diversification under specific settings of K. Specifically, we aim to examine the answers to the following questions: 1) are the differences between the diversified result and the baseline significant? and 2) are the differences between the LapPLSA regularized by subtopics from different external resources and the non regularized pLSA significant? For the first question, we consider the setting when K is set to the optimal value for a given topic set. For the second question, since both pLSA and LapPLSA are sensitive to K, we set K to the optimal value for each query, so that the impact of K is reduced to the minimum. Note that here by saying “optimal", we mean the best K in the range of $[2, 10]$ in terms of α-nDCG@20.

Table 4 and Table 5 show the results with respect to the first and the second questions, respectively.

From Table 4 we see that with a proper setting of K, in most cases, diversification with pLSA as well as with LapPLSA significantly outperforms the baseline. Again, the effectiveness of each individual resource differs for the three topic sets. For topics 1-50, regularization with click logs is most effective, while regularization with anchor text and with Web ngrams fails to have significant improvement over the baseline. On topic set 51-100, while pLSA based diversification does not result in significant improvement over the baseline in terms of both ERR-IA@20 and α-nDCG@20, all LapPLSA based runs show significant improvement in terms of at least one of the two measures. On topic set 101-150, only the run with LapPLSA using click logs does not achieve significant improvement over the baseline, all others do.

Table 5 shows that LapPLSA regularized with subtopics from external resources does not always lead to significant improvement over pLSA. However, on each of the three topic sets, at least one resource exists that, when used for regularization, outperforms pLSA significantly. Again, regularization with click logs is shown to be most effective, on topic set 1-50. On topic set 51-100, except click logs, regularization with one of the resources or their combinations results in significant improvement over the non-regularized pLSA.

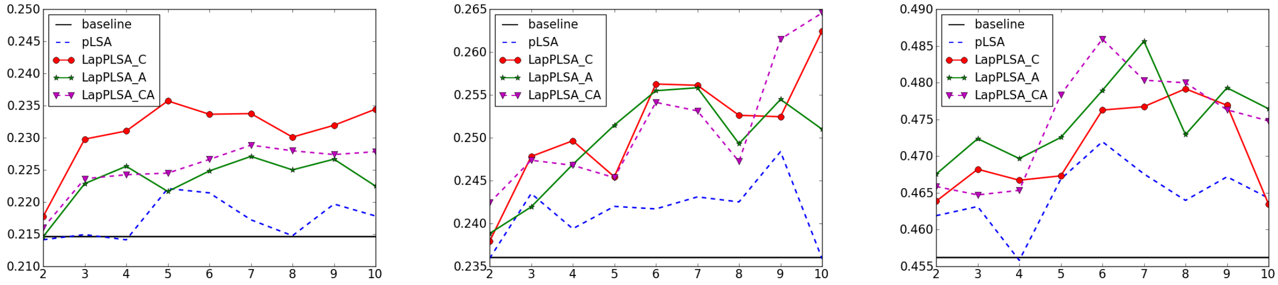


(a) Topics 1-50

(b) Topics 51-100

(c) Topics 101-150

Figure 3: Result of diversification in terms of α-nDCG@20 with topic models constructed using single resources. IA-select is used as the diversification method. The x-axis represents the value of K. The y-axis represents the α-nDCG@20 scores.



(a) Topics 1-50

(b) Topics 51-100

(c) Topics 101-150

Figure 4: Result of diversification in terms of α-nDCG@20 with topic models constructed using multiple resources. IA-select is used as the diversification method. The x-axis represents the value of K. The y-axis represents the α-nDCG@20 scores.

	Topics 1-50			Topics 51-100			Topics 101-150		
Method	K	E-IA@20	αnD@20	K	E-IA@20	αnD@20	K	E-IA@20	αnD@20
NoDiv	–	.130	.215	–	.161	.246	–	.354	.456
pLSA	5	.136$^{\triangle}$	.222$^{\triangle}$	9	.175	.259	6	.365$^{\blacktriangle}$	.472$^{\triangle}$
G^{C}	5	**.149**$^{\blacktriangle}$	**.236**$^{\blacktriangle}$	10	.184$^{\triangle}$	.273	8	.382	.479
G^{A}	7	.138	.227	7	.182	.266$^{\triangle}$	7	**.386**$^{\blacktriangle}$	**.486**$^{\blacktriangle}$
G^{N}	5	.138	.225	10	.186$^{\triangle}$	.272$^{\blacktriangle}$	7	.379$^{\triangle}$	.478$^{\triangle}$
G^{CA}	7	.139$^{\triangle}$	.229$^{\triangle}$	10	.186$^{\blacktriangle}$	**.276**$^{\blacktriangle}$	6	.385$^{\blacktriangle}$	**.486**$^{\blacktriangle}$
G^{CN}	10	.145	.228	7	.182$^{\blacktriangle}$	.267$^{\blacktriangle}$	6	.377$^{\blacktriangle}$	.481$^{\blacktriangle}$
G^{AN}	10	.138$^{\triangle}$	.224	10	**.187**$^{\triangle}$	.270$^{\triangle}$	7	.376$^{\blacktriangle}$	.477$^{\triangle}$
G^{CAN}	7	.135	.224	10	**.187**$^{\triangle}$	.274$^{\triangle}$	8	**.386**	.482$^{\triangle}$

Table 4: Diversification result with pLSA and LapPLSA regularized by different external resources and their combinations. All runs are compared to the baseline NoDiv. Boldface indicates the highest score among all runs.

	Topics 1-50		Topics 51-100		Topics 101-150	
Method	E-IA@20	αnD@20	E-IA@20	αnD@20	E-IA@20	αnD@20
pLSA	.149	.234	.193	.276	.397	.499
G^{C}	**.164**$^{\blacktriangle}$	**.257**$^{\blacktriangle}$	.201	.293	.410	.511
G^{A}	.147	.240	.200	.290$^{\triangle}$	.410	.509
G^{N}	.145	.235	.201$^{\triangle}$	.287$^{\blacktriangle}$	.409	.507
G^{CA}	.151	.244	**.207**$^{\blacktriangle}$	**.299**$^{\blacktriangle}$	.410	.510
G^{CN}	.153	.241	.200	.289$^{\triangle}$	.402$^{\triangle}$	.505$^{\triangle}$
G^{AN}	.144	.234	.196	.283$^{\triangle}$	.395	.499
G^{CAN}	.148	.240$^{\triangle}$	.198$^{\blacktriangle}$	.289$^{\blacktriangle}$	**.413**	**.512**

Table 5: Comparing LapPLSA and pLSA. All runs are compared to pLSA. All the scores are significantly greater compared to the baseline NoDiv in Table 4. Boldface indicates the highest score among all runs.

In fact, the performance of regularization with click logs is still decent; testing for significance of the difference between run G^{C} and run pLSA has a p-value of 0.077 for ERR-IA@20 and 0.059 for α-nDCG@20. The TREC 2011 topic set seems the most difficult one. Regularization with most resources or their combinations does not lead to significant improvement over the pLSA run. The only exception is the combination of the click logs and the Web ngrams. This result is to some extent consistent with statement in the TREC Web Track guideline that the topic set "*introduces "tougher" topics, ...they can rely less on click/anchor information, and popularity signals like PageRank.*"[14]. Combining all three resources seems to be a relatively safe choice: it improves significantly over the pLSA run on two out of the three topic sets, and on the third topic set, although the difference is not statistically significant (with a p-value of 0.1 for ERR-IA@20 and 0.054 for α-nDCG@20), the highest absolute score is achieved across all settings on this set.

6.4 A robustness analysis on the parameter K

From previous experiments, we have seen that the number of topics K is an important parameter, whose optimal value is difficult to predict. Further, we also see in Figure 3 and Figure 4 that across different settings of K, in most cases the averaged performance of LapPLSA exceeds that of pLSA. Given this observation, we are interested in the question: is regularized pLSA likely to outperform non-regularized pLSA no matter the value of K we select?

The above question can be reformulated as follows. Assume we have two samples of diversification results in terms of α-nDCG@20. Sample 1 is the result of diversification using pLSA for varying K, and sample 2 is the result of diversification using LapPLSA regu-

Resource	1-50		51-100		101-150	
	W	p-value	W	p-value	W	p-value
G^C	122	.0005	117	.0040	107	.0625
G^A	118	.0028	113	.0141	123	.0003
G^N	112	.0188	101	.1902	109	.0400
G^{CA}	**121**	.0007	**118**	.0028	**114**	.0106
G^{CN}	**116**	.0056	**109**	.0400	**118**	.0028
G^{AN}	108	.0503	**110**	.0314	**113**	.0142
G^{CAN}	109	.0400	**112**	.0188	**119**	.0019

Table 6: Comparing performance of LapPLSA and pLSA over random K's. Boldface indicates that the W value of a combined resource is equal or above the lowest W of the single resources that are combined.

larized by certain external subtopic resource, also for varying K. If we randomly pick a score from each sample, how probable does the score from sample 2 exceed the score from sample 1? This can be tested with a Wilcoxon ranksum test [40][4].

Table 6 shows the result. Each sample contains 9 observations, i.e., for $K = 2, ..., 10$. W is the rank sum statistics , where a larger W indicates a more extreme difference between the two samples. Two observations stand out. First, in all cases but three(G^{AN} on topics 1-50, G^N on topic 51-100, and G^C on 101-150), the differences between pLSA and LapPLSA are significant with a p-value < 0.05. That is, with a random setting of K, LapPLSA regularized with external resources tends to outperform non-regularized pLSA. Second, in most cases, the W value of those combined resources are in between (occasionally above) the resources that are combined. This is consistent with the observation made in Section 6.2. Further, compared to G^C and G^A, G^N has a relatively lower W on all three topic sets, which suggests that with a random K, LapPLSA regularized with G^N is less likely to improve over pLSA compared to G^A and G^C.

6.5 Summary

We conclude the experimental analysis by relating our findings to the research questions formulated before (See Section 5.1).

With respect to RQ1, we find that LapPLSA regularized with explicit subtopics extracted from good resources improves diversification results, which indicates that better topic models are formed.

With respect to RQ2, we find that different resources are effective on different topic sets. Futher, based on the observation in Figure 3 and the results discussed in Section 6.4, we conclude that the effectiveness of Web ngrams is the least robust, or more sensitive to the setting of K compared to anchor text and click logs. As mentioned before, we suspect that the Web ngrams are more similar to the document content than truly external resources, which could explain the observed difference.

For RQ3, we find that combining resources does not always improve the diversification result. The combined resource usually results in a diversification performance in between that of the individual resources combined.

In terms of RQ4, we find that LapPLSA regularized with explicit subtopics tends to outperform the non-regularized pLSA for cases where we do not optimize the setting of K, and simply choose it at random from a reasonable range. We therefore conclude that regularized pLSA has the advantage that it provides a more robust performance in practice (where we will not know the optimal K).

Topics			1-50				51-100				101-150	
Method	K	λ	E-IA@20	αnD@20	K	λ	E-IA@20	αnD@20	K	λ	E-IA@20	αnD@20
pLSA	5	.4	.136	.223	3	.1	.179	.258	5	.5	.368	.471
G^C	10	.1	**.167**	**.256**	10	.1	.195	.286	8	.2	.393	.486
G^A	3	.1	.145	.233	9	.1	.195	.286	7	.3	.396	.490
G^N	9	.1	.154	.240	10	.4	.184	.271	10	.3	.384	.482
G^{CA}	5	.1	.145	.240	7	.1	**.202**	**.295**	9	.3	**.397**	**.491**
G^{CN}	7	.1	.152	.243	7	.2	.187	.277	10	.3	.385	.485
G^{AN}	5	.1	.142	.233	10	.1	.192	.283	10	.2	.390	.485
G^{CAN}	3	.1	.148	.236	10	.2	.188	.279	8	.2	.392	.487

Table 7: Performance of xQuAD with pLSA and LapPLSA regularized by different external resources and their combinations. λ and K are optimized with respect to each topic set.

Finally, for completeness, we include the performance of xQuAD in Table 7. With a proper setting of λ and K, xQuAD shows better diversification performance compared to the results of IA-select in Table 4 in most cases. Same as with IA-select, regularization helps in generating better diversification results with xQuAD. Moreover, the usefulness of different resources and their combinations show similar trend compared to that of IA-select.

7. CONCLUSION

We proposed Multi-source Subtopics (MSS), a framework for subtopic modeling that i) uses a random walk based approach to combine and estimate the similarity between subtopics extracted from multiple Web sources, and ii) uses the obtained similarity relations to regularize topic models constructed using document content. MSS combines subtopics from multiple resources transparently, unifying implicit and explicit representations of subtopics.

We have demonstrated how the application of this framework in the context of search result diversification allows us to flexibly combine, analyze and compare subtopics extracted from different resources. Empirical results show that topic models regularized by the topical information extracted from external resources lead to improved and more robust diversification results, supporting our claim that better topic models are formed. Among the Web resources employed in our experiments, anchor text and click logs were shown to be generally more effective than Web ngrams. However, the effectiveness of a resource also depends on the properties of a specific query, and especially whether a resource contains subtopic information related to a query at all. Combining multiple resources could alleviate lack of coverage, but in our current setup leads to a diversification performance in between that of the resources that are combined.

A number of directions are left to be explored in the future. First of all, there is room to improve our method by exploring how to weigh resources by their (expected) quality for a given query. More resources can be analyzed, for example derived from social bookmarking sites, and more sophisticated ways of parameter optimization should be considered. Second, a component wise evaluation may improve our understanding beyond observations on the diversification pipeline as a whole. Hereto, we need to develop a method to evaluate the quality of the constructed subtopics directly through human assessment. This may be especially useful to improve our understanding of the features that make an external resource a good one for a given query. Finally, perhaps the strength of the external resources is not yet exploited in full when we only use them to regularize pLSA. An alternative could be to integrate implicit topics as nodes in the multi-plane random walks, and thus treat both types of topic representations more equally.

[4] This includes the assumption that diversification results are independent from each other with respect to different K's. We believe this assumption is sensible, especially given the observation that there is no clear pattern on the change of diversification performance with respect to the change of K.

Acknowledgements

This research was funded by European Commission FP7 grant 257024, in the Fish4Knowledge project

8. REFERENCES

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *WSDM '09*, pages 5–14, 2009.

[2] J. Allan and H. Raghavan. Using part-of-speech patterns to reduce query ambiguity. In *SIGIR '02*, pages 307–314, 2002.

[3] D. Beeferman and A. Berger. Agglomerative clustering of a search engine query log. In *KDD'00*, pages 407–416, 2000.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

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

[6] D. Cai, Q. Mei, J. Han, and C. Zhai. Modeling hidden topics on document manifold. In *CIKM '08*, pages 911–920, 2008.

[7] J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR '98*, pages 335–336, 1998.

[8] B. Carterette and P. Chandar. Probabilistic models of ranking novel documents for faceted topic retrieval. In *CIKM'09*, pages 1287–1296, 2009.

[9] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *CIKM'09*, pages 621–630, 2009.

[10] H. Chen and D. R. Karger. Less is more: probabilistic models for retrieving fewer relevant documents. In *SIGIR '06*, pages 429–436, 2006.

[11] C. Clarke, M. Kolla, G. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *SIGIR'08*, pages 659–666, 2008.

[12] C. Clarke, N. Craswell, and I. Soboroff. Overview of the TREC 2009 web track. In *TREC'09*, 2009.

[13] C. Clarke, N. Craswell, and I. Soboroff. Overview of the TREC 2010 web track. In *TREC'10*, 2010.

[14] C. Clarke, N. Craswell, and E. Soboroff, I.and Voorhees. Overview of the TREC 2011 web track. In *TREC'11*, 2011.

[15] G. V. Cormack, M. D. Smucker, and C. L. A. Clarke. Efficient and effective spam filtering and re-ranking for large web datasets. *Information Retrieval*, 14(5):441–465, 2011.

[16] N. Craswell and M. Szummer. Random walks on the click graph. In *SIGIR'07*, pages 239–246, 2007.

[17] N. Craswell, R. Jones, G. Dupret, and E. Viegas, editors. *WSCD'09*, 2009.

[18] J. Cui, H. Liu, J. Yan, L. Ji, R. Jin, J. He, Y. Gu, Z. Chen, and X. Du. Multi-view random walk framework for search task discovery from click-through log. In *CIKM '11*, pages 135–140, 2011.

[19] V. Dang and B. W. Croft. Query reformulation using anchor text. In *WSDM '10*, pages 41–50, 2010.

[20] V. Dang, X. Xue, and B. Croft. Inferring query aspects from reformulations using clustering. In *CIKM '11*, 2011.

[21] Z. Dou, S. Hu, K. Chen, R. Song, and J.-R. Wen. Multi-dimensional search result diversification. In *WSDM'11*, pages 475–484, 2011.

[22] A. Fuxman, P. Tsaparas, K. Achan, and R. Agrawal. Using the wisdom of the crowds for keyword generation. In *WWW '08*, pages 61–70, 2008.

[23] W. Goffman. A searching procedure for information retrieval. *Information Storage and Retrieval*, 2(2):73–78, 1964.

[24] J. Guo, X. Cheng, G. Xu, and X. Zhu. Intent-aware query similarity. In *CIKM'11*, pages 259–268, 2011.

[25] J. He. *Exploring topic structure: Coherence, Diversity and Relatedness*. PhD thesis, University of Amsterdam, 2011.

[26] J. He, E. Meij, and M. de Rijke. Result diversification based on query-specific cluster ranking. *J. Am. Soc. Inf. Sci. Technol.*, 62(3):550–571, 2011.

[27] D. Hiemstra and C. Hauff. MIREX: MapReduce information retrieval experiments. Technical Report TR-CTIT-10-15, University of Twente, 2010.

[28] T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR'99*, pages 50–57, 1999.

[29] Z. Li, F. Chen, Q. Xing, J. Miao, Y. Xue, T. Zhu, B. Zhou, R. Cen, Y. Liu, M. Zhang, Y. Jin, and S. Ma. Thuir at trec 2009 web track: Finding relevant and diverse results for large scale web search. In *TREC*, 2009.

[30] H. Ma, M. R. Lyu, and I. King. Diversifying query suggestion results. In *AAAI'10*, 2010.

[31] D. Metzler and W. B. Croft. Combining the language model and inference network approaches to retrieval. *Inf. Process. Manage.*, 40:735–750, September 2004.

[32] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *SIGIR '98*, pages 275–281, 1998.

[33] F. Radlinski, M. Szummer, and N. Craswell. Inferring query intent from reformulations and clicks. In *WWW '10*, pages 1171–1172, 2010.

[34] D. Rafiei, K. Bharat, and A. Shukla. Diversifying web search results. In *WWW '10*, pages 781–790, 2010.

[35] R. L. T. Santos, C. Macdonald, and I. Ounis. Exploiting query reformulations for web search result diversification. In *WWW'10*, pages 881–890, 2010.

[36] N. Tishby and N. Slonim. Data clustering by markovian relaxation and the information bottleneck method. In *NIPS*, pages 640–646, 2000.

[37] J. Wang and J. Zhu. Portfolio theory of information retrieval. In *SIGIR '09*, pages 115–122, 2009.

[38] K. Wang, C. Thrasher, E. Viegas, X. Li, and B.-j. P. Hsu. An overview of microsoft web n-gram corpus and applications. In *NAACL HLT'10*, pages 45–48, 2010.

[39] J.-R. Wen, J.-Y. Nie, and H.-J. Zhang. Clustering user queries of a search engine. In *WWW '01*, pages 162–168, 2001.

[40] F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

[41] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, 1st edition, 1999.

[42] Y. Yue and T. Joachims. Predicting diverse subsets using structural SVMs. In *ICML '08*, pages 1224–1231, 2008.

[43] C. X. Zhai, W. W. Cohen, and J. Lafferty. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *SIGIR '03*, pages 10–17, 2003.

[44] X. Zhu, A. B. Goldberg, J. Van, and G. D. Andrzejewski. Improving diversity in ranking using absorbing random walks. Technical report, University of Washington, 2007.