Constructing a Test Collection with Multi-Intent Queries

Ruihua Song[1,2], Dongjie Qi[3], Hua Liu[4], Tetsuya Sakai[2],
Jian-Yun Nie[5], Hsiao-Wuen Hon[2], Yong Yu[1]
[1]Shanghai Jiao Tong University, [2]Microsoft Research Asia, [3]Dalian University of Technology
[4]The Hong Kong University of Science and Technology, [5]University of Montreal
rsong@microsoft.com

ABSTRACT

Users often issue vague queries; when we cannot predict their intents precisely, a natural solution is to diversify the search results, hoping that some of the results correspond to the intent: This is usually called "result diversification". Only a few studies have been completed to systematically evaluate approaches on result diversity. Some questions still remain unanswered: 1) As we cannot exhaustively list all intents in an evaluation, how does an incomplete intent set influence evaluation results? 2) Intents are not equally popular; so how can we estimate the probability of each intent? In this paper, we address these questions in building up a test collection for multi-intent queries. The labeling tool that we have developed allows assessors to add new intents while performing relevance assessments. Thus, we can investigate the influence of an incomplete intent set through experiments. Moreover, we propose two simple methods to estimate the probabilities of the underlying intents. Experimental results indicate that the evaluation results are different if we take the probabilities into consideration.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Experimentation

Keywords

test collection, evaluation, diversity, ambiguous queries

1. INTRODUCTION

Queries issued by Web users often have multiple meanings or intents. Consider, for example, the keyword "TREC". This may refer to the Text Retrieval Conference, the Texas Real Estate Commission, or the equestrian sport of TREC. Without further information to disambiguate the intent of user, it is important for search engines to retrieve a set of diversified documents covering different requirements. Also it is better to rank the documents that are relevant to more popular intents higher than those relevant to less popular intents. Ideally, the document set should properly account for the interest of the overall user population [8].

EVIA 2010 June 15, Tokyo, Japan.

Sanderson [16] has surveyed previous research work on ambiguity and the effort taken to diversify search results. Although there is a long history of research addressing ranking problems for ambiguous queries, a lack of research work done to build test collections has hampered research of this type. This motivated TREC (Text Retrieval Conference) to evaluate search result diversity for the first time in 2009 [7]. In building the test collection, organizers chose to create subtopics (or intents) based on query log mining results before the procedure of assessing relevance, and regarded subtopics as equally important in calculating the Intent Aware metrics proposed by Agrawal et al.[1].

However, two important problems that may influence the effectiveness of test collections for evaluating search result diversity have not been discussed: 1) Exhaustively listing all intents of a given query may be impossible, but it is unknown how missing some intents will influence evaluation results; 2) It is reasonable to weight different intents by their popularity in measuring search result diversity, but few studies have been done to estimate the probability of each intent given a query.

We aim to address these two problems in this paper. First, we try to construct a test collection that has 50 multi-intent topics extracted from Wikipedia[1] disambiguation pages, e.g. the page for TREC[2], and a range of relevance judgments with regard to more than one interpretation. In addition to the initial intents extracted from Wikipedia, we provide assessors with a way to create new intents when they are assessing the relevance of documents. As a result, we can identify more intents. For example, for the query "TREC", the second most popular intent "Tennessee Real Estate Commission" is actually found during assessing relevance. In addition, we propose two methods to estimate the probability of an intent being underlying a given query. One is the log-based method and the other is the collection-based method. In the log-based method, we involve human assessors to judge the relevance of clicked URLs with regard to intents, and count the number of clicks associated with each intent to estimate the probabilities. In the collection-based method, we create sub-queries to retrieve the documents for different intents from the index of a Web collection, and count the number of returned documents in estimating the probabilities of intents.

The experimental results indicate that an incomplete intent set is less capable of discriminating some different rankings in evaluating diversity than a more complete intent set.

[1]www.wikipedia.org
[2]http://en.wikipedia.org/wiki/TREC

The absence of some popular intents also lowers the accuracy of the estimated probabilities of intents. Whether we apply the estimated probabilities or not does influence the evaluation results and the orders of simulated search results.

The remainder of this paper is organized as follows. In Section 2, we review related works. We describe our proposed approaches in Section 3 and conduct experiments in Section 4. Finally, we conclude and discuss future works in Section 5.

2. RELATED WORK

2.1 Diversification Algorithms

In many cases, users prefer a list of diverse results to a list of similar or near-duplicated results. Previous works have investigated result diversification in various applications, such as recommendation [26, 4], online shopping [19], query suggestion [18], personalized search [14], mining top stories in the blogosphere [11], and image retrieval [17]. Recently, the problem of improving search result diversity has attracted much interest [13]. To diversify search results, the probability of relevance of a document is assumed to be conditioned on the documents that appear before it in the result list. Carbonell and Goldstein [3] presented the Maximal Marginal Relevance (MMR) algorithm to diversify search results or document summaries. The method makes a tradeoff between novelty, measured by the similarity among documents, and the relevance, measured by the similarity between document and query. A parameter controls the degree of tradeoff. Agrawal et al. [1] proposed a greedy algorithm named IASelect to approximate the objective of minimizing the risk of dissatisfaction of the average user, in the setting where both queries and documents may belong to more than one category according to an existing taxonomy, such as the ODP taxonomy (`www.dmoz.org`). Some other researchers have proposed different diversification approaches, e.g. [21] based on structural SVMs, [24, 25] based on graph models, [23] based on a general risk minimization framework, and [6, 10]. As Sanderson noted in [16], "without test collections containing ambiguous topics with associated relevance judgments that reflect a range of interpretations of that topic, the worth of much of the work described here may not be fully understood." Therefore, we focus on the problems on building a test collection for evaluating diversification algorithms.

2.2 Evaluation Metrics on Diversity

Several metrics have been proposed to evaluate result diversity [5, 6, 8, 9, 15, 22]. Zhai, Cohen and Lafferty proposed a simple metric called S-recall (subtopic recall) in their proposed framework of subtopic retrieval. At document cutoff K, S-recall is defined as the number of unique subtopics that the top K results have covered divided by the number of all subtopics. Based on S-recall, they further defined S-precision and WS-precision.

Clarke et al. [8] proposed α-nDCG to evaluate search result diversity. For each document, they defined novelty-biased gain $NG(r)$ as follows:

$$NG(r) = \sum_{i=1}^{m} J_i(r)(1-\alpha)^{C_i(r-1)}$$

where, $C_i(r-1)$ is the number of relevant documents found within the top $r-1$ documents for intent i. $J_i(r)$ is a binary variable indicating whether the document at rank r is relevant to intent i or not. When α is closer to 1, the novelty is rewarded more in the metric. Clarke et al. proposed a new computation of nDCG [12] based on $NG(r)$.

Agrawal et al. [1] proposed a family of Intent Aware (IA) metrics. Let i be an intent and suppose that for each query q, the probabilities of different intents $p(i|q)$ are given. Then $MAP\text{-}IA$ is given by

$$MAP\text{-}IA = \sum_i p(i|q) MAP_i$$

Other metrics, such as $nDCG\text{-}IA$, can be defined in a similar way.

As we address the problem of estimating $p(i|q)$ in this paper, we use $MAP\text{-}IA$ as our main metric in experiments.

2.3 Test Collections on Evaluating Diversity

The TREC 2009 Web track[3] diversity task is the most relevant practice on constructing a test collection for evaluating diversity [7]. Different from the adhoc task, given a query, organizers extracted and analyzed groups of related queries, using co-clicks and other information, to identify clusters of queries that highlight different aspects and interpretations of the target query. The set of subtopics is not exhaustive with the number of subtopics per query ranging from three to eight, with a mean of 4.9. Documents were judged with respect to the subtopics. The judgments are binary as to whether or not the document satisfies the information need associated with the subtopic. In addition, the diversity task used new measures, i.e. Intent Aware precision and α-nDCG. They assume equal probabilities of different subtopics in calculating Intent Aware precision.

As a preliminary trial, the diversity task of TREC 2009 Web track provides valuable data and experiences on evaluating result diversity. Motivated by this work, we aim to address some problems that remain. On the one hand, the subtopics in TREC 2009 Web track diversity task were created before assessing relevance. When assessors find a document that is relevant to the query but its relevant subtopic is not covered by the pre-defined subtopics, there is no way to add a new subtopic. As a result, the documents relevant to some important but missing subtopics have to be judged as irrelevant. For example, for topic wt09-6 "kcs", King's College School is not included in the subtopics. On the other hand, for Intent Aware measurements, the likelihood of the intents is important to consider in evaluating diversified systems. Instead of using equal probabilities, we investigate the probability estimation methods.

Related to the probability estimation of intents, Agrawal et al. [1] set up a test collection to evaluate their proposed diversification algorithms. In their setting, the categories of queries can be viewed as implicit intents. They submitted queries along with the most likely three categories as estimated by a classifier to Amazon Mechanical Turk platform (`www.mturk.com`). For each query, seven Turks were asked to associate the query with the closest category. As different Turks selected different categories for 70% of the queries, they used the data to estimate $p(c|q)$. Different from this work, we estimate the probabilities for explicit intents, instead of the ODP categories.

[3]TREC 2009 Web Track: `http://plg.uwaterloo.ca/~trecweb`

3. OUR APPROACH

In this section, we describe how we construct a test collection of multi-intent queries, the intents of which can be incrementally added by assessors during the stage of judging relevance. For each intent, we propose two simple methods to estimate its probability.

3.1 Building a Test Collection

In general, an IR test collection is comprised of queries, documents, and judgments for query-document pairs. For multi-intent queries, the intents that a document is relevant to are also required for evaluating diversity.

It is challenging to sample multi-intent queries and enumerate their different intents. First, a set of multi-intent queries proposed by a few people tend to be biased by individual experiences. Second, it is costly to sample these kinds of queries from query logs manually because it is difficult for humans to judge whether a query has multiple intents due to limited knowledge. Third, even if we have multi-intent queries sampled, there are still difficulties in listing all the intents of a query. Fortunately, thousands of people contribute a huge amount of knowledge to Wikipedia. For an entry with multiple interpretations, Wikipedia provides a disambiguation page to allow users to choose interpretations that are of their interest. This resource is valuable for the identification of a set of multi-intent queries and their possible interpretations. In our experiments, we leverage Wikipedia to build a test collection for evaluating diversity.

We make use of disambiguation pages to identify multi-intent titles as Sanderson does in [16]. We also filter the titles from Wikipedia by checking whether a title is among the list of queries from search engine logs for half a year. This is to make sure that our sampled multi-intent entries are real web queries. Then we sample 50 representative queries, in which some queries have more diverse intents than others. For example, "TREC"[4] refers to Text Retrieval Conference, Texas Real Estate Commission, the Trans-Mediterranean Renewable Energy Cooperation, etc., which are totally unrelated entities. In contrast, "A Beautiful Mind"[5] tends to have more similar intents, such as A Beautiful Mind (book), A Beautiful Mind (film), and A Beautiful Mind (soundtrack).

We collect the top returned documents from two search engines to form a document set for assessing relevance. Two strategies can be applied in collecting documents. One is to use the multi-intent queries only to retrieve documents, and we denote this document set P. The other strategy is to use both the multi-intent queries, e.g. "A Beautiful Mind", and the sub-queries that correspond to intents, e.g. "A Beautiful Mind book", "A Beautiful Mind film", and "A Beautiful Mind soundtrack", to retrieve documents, and we call this document set as P^+. We choose the second strategy in our experiments because P may not cover some unpopular meanings because the corresponding documents are missing from the retrieval results. For example, for a query with a dominating intent, such as "Java", its top search results may only cover the intent of "Java programming", although it may also refer to "island of Java". Thus, we submit the query and its additional sub-queries respectively to two commercial search engines and retrieve the top 20 returned documents for each query/sub-query. By merging the retrieved documents and removing duplicates, we make a pool of documents for assessing relevance.

[4] http://en.wikipedia.org/wiki/TREC

[5] http://en.wikipedia.org/wiki/A_beautiful_mind

Table 1: Intents of "TREC" (The initial intent set I is composed of the first four intents)

No.	Intent
1	Text Retrieval Conference
2	Texas Real Estate Commission
3	Trans-Mediterranean Renewable Energy Co.
4	T-cell receptor excision circles
5	Tennessee Real Estate Commission
6	TREC-UK sport of TREC
7	trec horse rider
8	Tropical Research Education Center
9	T-Cell Rearrangement Excision Circle
10	Text Retrieval and Evaluation Conference
11	TREC educational research experience
12	BHS TREC sport
13	Tom Ridge Environmental Center
14	Tutoring Reading Enabling Children
15	Toronto Renewable Energy Co-operative
16	Training Resources Environmental Community
17	trec management mining oil
18	Tallapoosa River Electric Cooperative
19	Tissue Repair Engineering Centre
20	Transition Resources Education Community
21	TREC low-stress competition
22	Transdisciplinary Research Energetic Cancer
23	TREC randomized pragmatic trial
24	Tenderloin Reflection Education Center

We developed a labeling tool to judge whether a document is relevant to a query as well as which main intents the document covers. As Figure 1 shows, the frame on the right hand side displays the page with keywords highlighted. On the left questionnaire frame, an assessor can mark a page as "Not Found", if the page fails to load; or "Irrelevant", which means the page's content is not relevant to the query at all; or "Relevant", which means the page content is relevant to the query. If "Relevant" is checked, the assessor is also asked to choose one or more relevant intents from a list of candidates. If the assessor finds a new intent, he/she could add this to the candidate list through the text box at the bottom left. At the beginning, the intents shown in the interface are extracted from Wikipedia, which composes an initial intent set, denoted as I; when the assessor finishes assessing all documents, we have an expanded set of intents, denoted as I^+. For instance, only four intents are extracted from the Wikipedia disambiguation page, whereas 20 more intents are added during the assessments, as shown in Table 1. Thus, I contains the first four intents, and I^+ contains all the 24 intents.

Finally, in the judgments, we have the following information for a given query:

1. All intents, including Intent IDs and descriptions;
2. A relevance tag, i.e. Not-Found/Irrelevant/Relevant, assigned to each document;
3. A set of relevant Intent IDs associated with a relevant document.

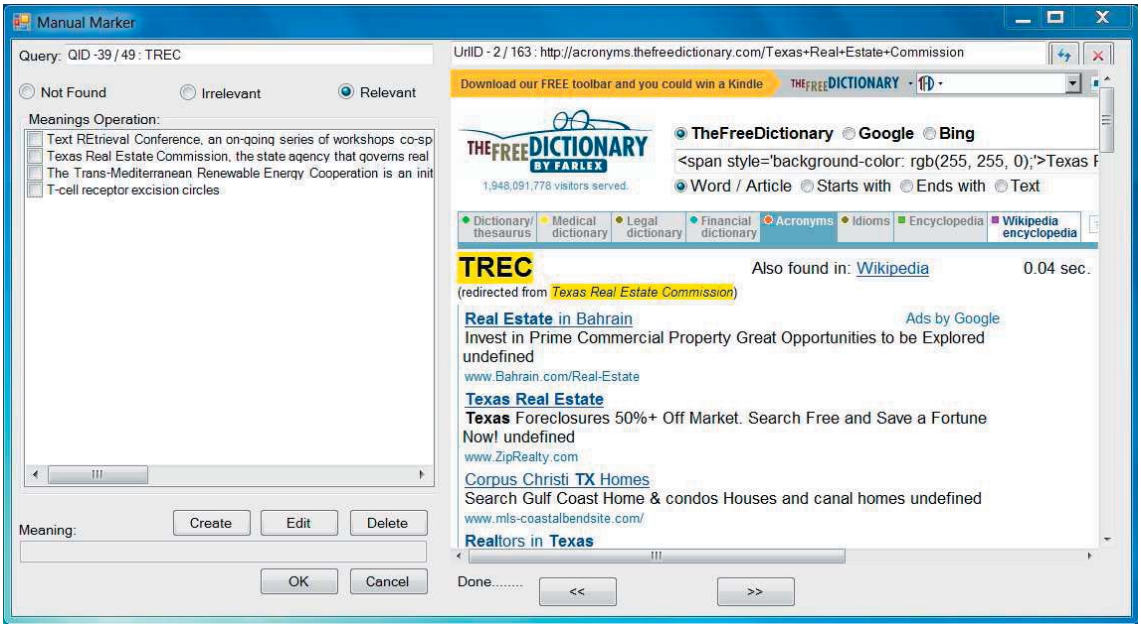


Figure 1: Illustration of the labeling tool

3.2 Estimating $p(i|q)$

The probability that an intent i underlies a given query q is used to weight intents in diversity measures such as Intent Aware metrics proposed by Agrawal et al. [1]. In this section, we propose two methods to estimate $p(i|q)$.

3.2.1 A Log-based Method

From the perspective of search engine users, we propose using search logs to estimate the probability $p(i|q)$, which reflects how popular an intent is. Click-through logs record the queries that a user issued and the corresponding URLs that the user clicked. By issuing the same query, different users may have distinct intents, thus click on different documents. Here, the clicked documents can provide us with the valuable information on which intents the user has when issuing the query. Therefore, we propose estimating $p(i|q)$ in three steps:

First, we ask assessors to associate each clicked document with relevant intents. The labeling tool described in Section 3.1 is applied. As a result, we know whether a clicked document d is relevant to an intent i. If yes, $rel(d,i) = 1$; otherwise, $rel(d,i) = 0$.

Second, we count the number of clicks for each intent:

$$count(i,q) = \sum_{rel(d,i)=1, d \in C} click(d)$$

Here, C is the set of documents that have been clicked for the query q; $click(d)$ is the number of aggregated times that users have clicked d.

Finally, we calculate $p(i|q)$ following the maximum likelihood estimation with Laplace smoothing (a.k.a. Add-One smoothing):

$$p(i|q) = \frac{count(i,q)+1}{\sum_{i \in I^+}(count(i,q)+1)}$$

Maximum likelihood estimation works fine for data that occur frequently in the training corpus, but when an intent does not occur, it does not mean that it should have probability zero due to the sparse data problem. By multiplying the zero $p(i|q)$ to the measurement for the intent i, this unseen intent cannot influence the diversity metric no matter how the relevant documents are ranked. Thus, we add one to all counts to smooth the probabilities.

A possible problem with this basic idea is that some documents that are relevant to minor intents may be missing in the clicked document set C. For example, "Java programming" is the dominant intent for the query "java". Almost all the returned documents on the first few pages are relevant to this major intent. If we use the clicked document for "java" only, we will miss the clicked documents that are relevant to the minor intents "island of Java" and "Java coffee".

To avoid this problem, we propose using the follow-up queries q' that have been issued after the query q to expand the set of clicked documents C to C^+. Similar to [2], we construct session data by the following steps: 1) We identify each individual user's logs to obtain a separate stream of query events; 2) we separate two consecutive queries into two sessions if the time interval exceeds 30 minutes, which is a widely-used rule [20]; 3) we store each pair of two consecutive queries of the sequence of session into a hash table. Thus, we can easily find q' that has followed q. We add the clicked documents for all q' to C and obtain the expanded set C^+. Hopefully, more relevant documents to the minor intents could be included in C^+.

As all the clicked URLs are assessed manually, unrelated follow-up queries cannot hurt the probability estimation. Based on the clicks of relevant URLs in C^+, we can count the number of clicks for each intent and estimate $p(i|q)$ in the same way.

3.2.2 A Collection-based Method

From the perspective of Web publishers, we assume that the more interest a topic attracts, the more documents on

the topic would be published on the Web. Thus, we propose using web collection to estimate the probability $p(i|q)$. This is estimated as follows.

First, for any $i \in I^+$, we could create a sub-query that is able to distinguish this intent from others. Meanwhile, the sub-query is as short as possible. For instance, Table 1 shows the sub-queries that we created for different intents of TREC. We assume that the documents that contain all the terms of a sub-query are relevant to the corresponding intent i. In the example of "TREC", the document containing "text", "retrieval", and "conference" is assumed relevant to the intent of the "TREC" conference, whereas the document containing "texas", "real", "estate", and "commission" is assumed relevant to the intent on the real estate organization in Texas.

Second, we index all English pages in ClueWeb09[6]; there are about five million English pages in the collection. We do not filter any stop words, nor perform stemming in parsing documents and queries. When submitting a query, we retrieve the document lists for individual query terms from the index, and join the lists to return the number of documents containing all query terms, denoted as N_i. For example, 71,405 documents contain "text", "retrieval", and "conference", whereas 238,175 documents contain "texas", "real", "estate", and "commission".

Finally, we estimate the probability $p(i|q)$ by the following formula:

$$p(i|q) = \frac{N_i + 1}{\sum_{i \in I^+} (N_i + 1)}$$

Again, we use Laplace smoothing to solve the problem on the intents with zero N_i.

Notice that the above estimation is only an approximation. In particular, we assume that the number of returned documents containing all the terms in a sub-query are relevant to the corresponding intent, which could turn out wrong. In a later section, we will analyze this strategy in comparison to the log-based estimation.

4. EXPERIMENTS

We set up a test collection of 50 queries and perform a series of experiments to investigate the factors that impact on the evaluation of search result diversity. In our experiments, seven assessors, comprising six females and one male, are hired to judge the relevance of documents. It takes 543 hours in assessing the relevance of documents for evaluation, whereas it takes 529 hours in assessing the relevance of clicked documents for estimating probabilities. All assessors are undergraduate or graduate students, who are major in Materials, Economic Laws, Linguistics, and Computer Science. They are fluent in English and familiar with Web search. Each assessor is assigned a few queries, and then he/she is responsible for labeling all documents on the queries. Thus, the assessor could keep consistent standards in judging relevance and adding new intents.

4.1 Statistics on the Built Test Collection

By applying the methodology described in Section 3.1, we sampled 50 multi-intent queries from Wikipedia and built a test collection for evaluating diversity. Some important statistics are shown in Table 2.

[6]ClueWeb09: `http://boston.lti.cs.cmu.edu/Data/clueweb09`

Table 2: Statistics on the test collection containing 50 multi-intent queries from Wikipedia

DSet	Numbers per query	[Min, Max]	Mean
	#Intents(I)	[1, 14]	5.8
P	#Docs judged	[29, 40]	33.74
	#Relevant docs	[2, 32]	18.7
	#Intents(I^+)	[2, 32]	10.58
P^+	#Docs judged	[50, 540]	247.8
	#Relevant docs	[6, 318]	147.94
	#Intents(I^+)	[2, 46]	15.82

Table 3: Comparing four implementations of the log-based method in estimating $p(i|q)$

Duration	Query Set	#Intents(clicked)	#Clicks
one-month	C	134	8206
one-month	C^+	183	35692
Improvements		36.6%	335.0%
four-month	C	141	24196
four-month	C^+	**167**	**98417**
Improvements		18.4%	306.7%

On average, 247.8 documents per query are judged by assessors, in which 60% of documents are judged as relevant to different intents. The number of initial intents ranges from 1 to 14 per query, with a mean of 5.8. As the labeling tool allows annotators to add new intents, the number of intents dramatically increases from 5.8 to 15.8 per query after labeling. This large increase shows that the initial set of intents is very limited and many possible intents are not considered. In the assessments, about 9 documents are relevant per intent.

In Section 3.1, we propose to expand the document set P by adding the top retrieved documents for the sub-queries corresponding to initial intents. The expanded document set is P^+. As Table 2 shows, if we judge the documents in P only, the number of documents per query decreases from about 248 to 34, where the number of relevant documents is reduced by 78%. Consequently, about 33% of intents will not be found. This indicates that expanding the document set for assessing relevance can improve the coverage of intents and thus potentially make the built test collection more complete.

In the remainder of this document, we use P^+ document set and corresponding judgments in evaluation.

4.2 Estimating $p(i|q)$

4.2.1 Experiments on Log-based Methods

There are two ways to alleviate the data sparseness of query logs: 1) We could increase the amount of log data in estimating the likelihood of intents; 2) as described in Section 3.2.1, we use user session data to expand the original query set C by the follow-up queries, and thus the query set used for extracting clicked URLs is enlarged to C^+. In our experiments, there is a one-month log in November, 2009 and a four-month log from July to October, 2009. We collect the clicked URLs from the one-month log for the expanded query set C^+, and ask assessors to associate the URLs with relevant intents. The assessments provide useful data to compare the four log-based methods with different durations or query sets. The results are shown in Table 3.

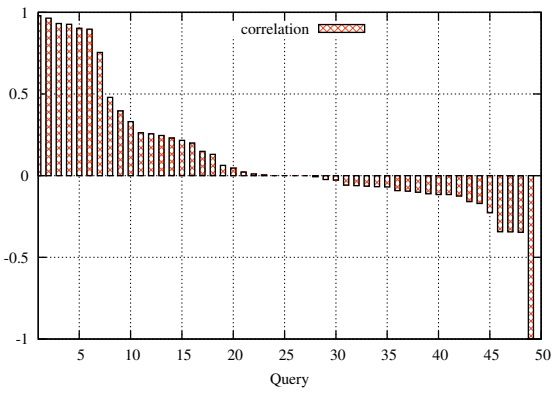


Figure 2: Correlation between $p(i|q)$ estimated by the log-based method and that estimated by the collection-based method

The results indicate that expanding queries using follow-up queries can increase the number of covered intents by 36.6% and the number of clicks by more than three times. The increase is also confirmed by the statistics on four-month logs. This suggests that the follow-up queries are effective in partially solving the sparseness problem.

In addition, the number of clicks over the four-month data increases about 1.75 times, compared to the number of clicks over the one-month data. This confirms that more query logs are effective in alleviating the data sparseness problem. Notice that the one-month data has no overlap with the four-month data. We observe that the number of covered intents with clicks drops a little bit in four-month data. This phenomenon is caused by time sensitiveness: 1) Some intents, e.g. Text Retrieval Conference, do not interest users constantly, and thus there is some difference between query logs from different periods of time; 2) some relevant URLs are changed, but not all the clicked URLs in the four-month log are judged. If we judge all clicked URLs over the four-month data, the cost will be also increased by 1.75 times. Therefore, there is a trade-off between the duration of logs and the labor cost on relevance assessments.

In the remaining experiments, the log-based method refers to the estimation results over the four-month logs and C^+, if not specified otherwise.

4.2.2 Log-based Method vs. Collection-based Method

As described in Section 3.2, given a query, we have a probability vector estimated by the log-based method and a probability vector estimated by the collection-based method. To compare these two methods, we calculate the correlation between two vectors for each of the 50 queries. For example, given a query q with n intents, the log-based method estimates a vector of $< p_1^l, p_2^l, \cdots, p_n^l >$, whereas the collection-based method outputs a vector of $< p_1^c, p_2^c, \cdots, p_n^c >$. Then we caculate the correlation between two vectors for the query q. The correlation coefficients for all queries are sorted in descending order and shown in Figure 2.

Surprisingly, we find that the correlation between the two methods is lower than 0.3 for about 80% of the queries. When looking into the probability vector for each query, we have the following observations:

- Figure 3 (a) shows the estimated probabilities for *trec*, which represents the queries whose intents can be easily differentiated by sub-queries. TREC is an acronym that has 24 meanings. The log-based methods and the collection-based method agree on the most popular two intents. However, the probability for "Text Retrieval Conference" estimated by logs is much lower than that estimated by the number of relevant documents. This shows a reasonable gap between web publishers and general web users, and the correlation is 0.96.
- Figure 3 (b) shows the estimated probabilities for *midweek*, which represents the queries whose intents cannot be easily differentiated by sub-queries. "Midweek" could be a magazine based in Hawaii, whereas its literal meaning is the middle of a week. However, it is difficult to create a sub-query that can distinguish this meaning from others. By the sub-query "midweek week", most likely some documents that are relevant to other meanings are also counted for this literal meaning. That is why this meaning gains the largest probability in the collection-based method. The correlation is -0.16. Similar confusion occurs for "urgent", "Laila" (a person's name), "Poinole" (a location name), etc.

Due to the above reason, although the collection-based method requires less labor, it is not applicable to all multi-intent queries due to the difficulty in differentiating intents by sub-queries. Therefore, we trust more the probabilities estimated from the log-based method, and use them in the following experiments.

4.3 Evaluating Diversity

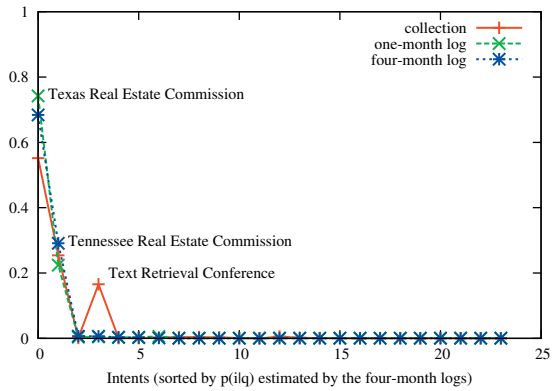
4.3.1 Evaluation Experiments on Intents

As described in Section 3.1, we allow assessors to add new intents that are not included in the initial intent set I extracted from Wikipedia. In this experiment, we take "TREC" as an example to show how the intents influence the evaluation of diversity.

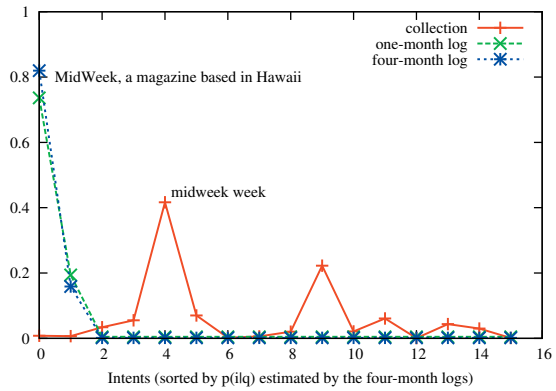
As shown in Table 1, during the assessment, assessors find 20 more intents for the abbreviation *TREC*. In terms of clicks, the added intent "Tennessee Real Estate Commission" attracts significantly more interests from users than "Text Retrieval Conference" (See Figure 3 (a)). This indicates that although Wikipedia contains a large amount of user contributed contents, the meanings of an entry are far from exhaustively covered and some important meanings may be missing. It is necessary to complete new and important intents by other means, such as through the interface we designed.

To see the impact of the completeness of identified intents on the evaluation of diversity, let us use I and I^+ respectively to evaluate two search engine results using the query "TREC" as shown in Table 4. Based on I^+, Search Engine 1 (SE1) returns the documents that cover three intents, i.e., i_1, i_2, and i_{15}, whereas, Search Engine 2 (SE2) returns the documents that are relevant to two intents only. The evaluation results are shown in Table 5.

According to the table, we have three observations on the influence of intent incompleteness. First, the estimated probabilities of intents are different. For example, without adding new intents, $p(i_2|q)$ is 0.9868, which is significantly larger than the probability (0.6840) estimated based on the expanded intent set I^+. Consequently, the Intent Aware



(a) trec



(b) midweek

Figure 3: Comparison of $p(i|q)$ estimated by different methods for two example topics

Table 4: Top five search results returned by two search engines for "TREC"

	Search Engine 1		Search Engine 2	
1	www.trec.state.tx.us	i_2	www.trec.state.tx.us	i_2
2	(irrelevant)	-	(irrelevant)	-
3	trec.nist.gov	i_1	trec.nist.gov	i_1
4	www.trec.on.ca	i_{15}	(irrelevant)	-
5	(irrelevant)	-	(irrelevant)	-

Table 5: Calculating MAP-IA@5 of two rankings by different $p(i|q)$ for "TREC"

I^+	$p(i\|q)$		MAP_i	
	Uni	Log	SE1	SE2
i_1	0.0417	0.0065	$1/3/44 = 0.0076$	$1/3/44 = 0.0076$
i_2	0.0417	0.6840	$1/1/30 = 0.0333$	$1/1/30 = 0.0333$
i_{15}	0.0417	0.0011	$1/4/1 = 0.25$	0
MAP-IA@5(U)			**0.0121**	0.0017
MAP-IA@5(L)			**0.0231**	0.0229
I	$p(i\|q)$		MAP_i	
	Uni	Log	SE1	SE2
i_1	0.25	0.0093	$1/3/44 = 0.0076$	$1/3/44 = 0.0076$
i_2	0.25	0.9868	$1/1/30 = 0.0333$	$1/1/30 = 0.0333$
MAP-IA@5(U)			0.0102	0.0102
MAP-IA@5(L)			0.0330	0.0330

metrics will more heavily rely on the major intent if some important intents are missing. Second, an incomplete intent set may not discriminate the search results with a difference in diversity. For example, by using I, we cannot identify the document *www.trec.on.ca* that is relevant to intent No.15. As a result, SE1 and SE2 perform the same in terms of MAP-IA, no matter whether equal probabilities or estimated probabilities are used. Third, when the number of relevant documents is too small for an intent i, MAP_i may play an over-important role in calculating $MAP\text{-}IA$. For example, as only one document is relevant to i_{15}, MAP_i is much larger than the other two intents when we are evaluating SE1. Although i_{15} is a minor intent, SE1 significantly outperforms SE2 in terms of $MAP\text{-}IA@5$ based on the uniform distributed probabilities.

Therefore, adding new intents can improve the accuracy of estimated probabilities, and discriminate more between ranking results in diversity, but we should pay attention to

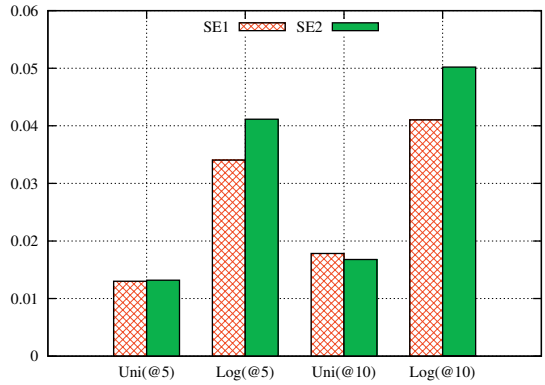


Figure 4: Evaluating two search engines in terms of MAP-IA by $p(i|q)$ of uniform distribution and $p(i|q)$ estimated by four-month logs

the effect of intents with very few relevant documents.

4.3.2 *Evaluation Experiments on* $p(i|q)$

We evaluate how two search engines perform in search result diversification using the test collection of 50 queries based on I^+. The mean performance results are shown in Figure 4. They are divided into four groups: 1) Uni(@5) is measured by $MAP\text{-}IA@5$ using the uniform $p(i|q)$; 2) Log(@5) is measured by $MAP\text{-}IA@5$ using the estimated $p(i|q)$ from logs; 3) Uni(@10) is measured by $MAP\text{-}IA@10$ using the uniform $p(i|q)$; 4) Log(@10) is the result measured by $MAP\text{-}IA@10$ using the estimated $p(i|q)$.

As shown in the figure, we find that the comparison results of SE1 and SE2 are not consistent when different probabilities of intents are applied. In terms of $MAP\text{-}IA@10$, SE1 slightly outperforms SE2 if we apply the uniform probabilities, whereas, SE2 performs significantly better than SE1 if we apply the estimated probabilities. Moreover, in terms of $MAP\text{-}IA@5$, the performance of SE1 is very close to that of SE2. However, when weighted by the estimated probabilities, $MAP\text{-}IA@5$ shows an obvious gap between two engines. This indicates that SE1 may return more results on minor intents at the top whereas SE2 does better in ranking the documents that are relevant to major intents. If we care

Table 6: Top five search results returned by two search engines for "midweek"

	Search Engine 1		Search Engine 2	
1	midweek.com	i_3	midweek.com	i_3
2	midweeknews.com	i_3	midweek...movies/	i_3
3	(irrelevant)	-	bbc.co.uk/...qrpf	i_2
4	(irrelevant)	-	(irrelevant)	-
5	en.wikipedia...Midweek	i_1	midweeknews.com	i_3

Table 7: Calculating MAP-IA@5 of two rankings by different $p(i|q)$ for "midweek"

	$p(i\|q)$		MAP_i	
	Uni	Log	SE1	SE2
i_1	0.0625	0.8197	0.04	0
i_2	0.0625	0.0015	0	0.0175
i_3	0.0625	0.0015	0.1818	0.2364
MAP-IA@5(U)			0.0139	**0.0159**
MAP-IA@5(L)			**0.0331**	0.0004

about the average satisfaction of users, the Intent Aware metrics with the estimated probabilities may be better than those with the uniform probabilities.

To better understand the impact of $p(i|q)$, we take "midweek" as an example. Table 6 lists the top five search results from the two search engines. For relevant documents, we also show which intent it is relevant to. The probabilities of intents and evaluation results are shown in Table 7. We find that SE2 outperforms SE1 in terms of the $MAP\text{-}IA$@5 with the uniform probabilities. This is somehow reasonable because SE2 returns one more relevant document than SE1, and both engines cover two intents of "midweek". However, in terms of user clicks, i_1 is much more intended by users than the other two intents are. As a result, SE1 outperforms SE2 in terms of $MAP\text{-}IA$@5 with the estimated probabilities from logs, because it returns one more document that is relevant to the most important intent i_1.

4.4 Ranking Search Results

We conduct simulation experiments to investigate how the test collection performs with different settings in ranking search results.

First, we generate 100 simulated runs. Given a query, we merge the top 20 URLs from SE1 and the top 20 URLs from SE2 and remove duplicated URLs. Now we have a list of unique URLs. Then, in each run, we shuffle the list three times to ensure that the order of URLs is randomized. Then we select the first ten URLs to form a search result for the query. Each run contains the simulated search results for 50 queries.

Second, we evaluate all runs and output $MAP\text{-}IA$@10 averaged over 50 queries. Four evaluation settings are applied respectively: 1) $I\text{-}U$: using the initial intent set I with the uniform probabilities; 2) $I\text{-}L$: using the initial intent set I with the estimated probabilities from logs; 3) $I^+\text{-}U$: using the expanded intent set I^+ with the uniform probabilities; 4) $I^+\text{-}L$: using the expanded intent set I^+ with the estimated probabilities from logs.

Finally, we calculate Kendall's τ between the lists of runs ordered by different evaluation settings. Kendall's τ is defined as:

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}$$

Table 8: Average Kendall's τ rank correlations between the lists of ordered 100 simulated runs with different evaluation settings

	$I\text{-}U$	$I\text{-}L$	$I^+\text{-}U$	$I^+\text{-}L$
$I\text{-}U$	1	0.415	0.179	0.224
$I\text{-}L$	-	1	0.087	0.425
$I^+\text{-}U$	-	-	1	0.147
$I^+\text{-}L$	-	-	-	1

where n_c is the number of concordant run pairs between two lists, n_d is the number of discordant run pairs, and n is the number of runs, e.g. 100 in our experiments.

To avoid accident, we conduct the experiment ten times and average the Kendall's τ for a given pair of evaluation settings. The experimental results are shown in Table 8.

The table shows that the correlation coefficients between different pairs of settings are low except for the pair of $I\text{-}U$ and $I\text{-}L$, and the pair of $I\text{-}L$ and $I^+\text{-}L$. If we use the initial intent set I, the $MAP\text{-}IA$@10 with the uniform probabilities is relatively similar to that with the estimated probabilities in ranking runs. If we use the log based method to estimate probabilities, the evaluation results based on I are relatively similar to those based on I^+ in ordering the simulated runs. However, the other low correlations indicate that substantial difference does exist between I and I^+ if the uniform probabilities are applied, and so does between the uniform probabilities and estimated probabilities over I^+. Therefore, we should be aware of the difference when we use incomplete intent sets or the equal probabilities in approximately calculating intent aware metrics.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we investigated some open questions on the practice of evaluating search result diversity. First, we allowed assessors to add new intents other than those predefined intents in judging relevance in order to obtain a more complete list of intents. Second, we proposed two simple methods to estimate the probabilities of intents being underlying a query, based on either query logs or the collection. Third, we conducted experiments to compare different methods of evaluating diversity in order to see the impact of incompleteness of the list of intents as well as the estimation of intent probabilities. Our experimental results showed that adding new intents can improve the accuracy of estimated probabilities, and discriminate more between ranking results in terms of diversity. Substantial difference between the sets with or without added intents is observed on simulation data if intents are regarded as equally popular. In addition, the log-based method is more reliable than the collection-based method because of the difficulties in creating sub-queries to distinguish intents in the collection-based method. The Intent Aware metric $MAP\text{-}IA$ is influenced by whether we use the estimated probabilities or not.

For future work, we will investigate the following questions. First, if a query is faceted, how could we control the granularity of added intents/subtopics? Second, how could we design a user interface to enable graded-relevance judgments per intent? Third, in this paper, we have only the search results from two search engines. In the future, we will try to involve real participants to retrieve and diversify search results in a static collection. Thus, we could verify our findings on real runs instead of simulated runs.

6. REFERENCES

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *WSDM '09: Proceedings of the 2nd ACM international conference on Web Search and Data Mining*, pages 5–14, 2009.

[2] H. Cao, D. Jiang, J. Pei, Q. He, Z. Liao, E. Chen, and H. Li. Context-aware query suggestion by mining click-through and session data. In *KDD '08: Proceedings of the 14th ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, 2008.

[3] J. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval*, 1998.

[4] O. Celma and P. Herrera. A new approach to evaluating novel recommendations. In *RecSys '08: Proceedings of the 2nd ACM international conference on Recommender Systems*, 2008.

[5] O. Chapelle, D. Metzler, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *CIKM '09: Proceedings of ACM 17th Conference on Information and Knowledge Management*, pages 621–630, 2009.

[6] H. Chen and D. R. Karger. Less is more. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval*, pages 429–436, 2006.

[7] C. L. A. Clarke, N. Craswell, and I. Soboroff. Overview of the trec 2009 web track. In *TREC2009: Proceedings of the 18th Text Retrieval Conference*, 2009.

[8] C. L. A. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Buttcher, and I. Mackinnon. Novelty and diversity in information retrieval evaluation. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval*, pages 659–666, 2008.

[9] C. L. A. Clarke, M. Kolla, and O. Vechtomova. An effectiveness measure for ambiguous and underspecified queries. In *Advances in Information Retrieval Theory (ICTIR 2009), LNCS 5766*, pages 188–199, 2009.

[10] M. Coyle and B. Smyth. On the importance of being diverse: analysing similarity and diversity in web search. In *Intelligent Information Processing II*, pages 341–350.

[11] K. El-Arini, G. Veda, D. Shahaf, and C. Guestrin. Turning down the noise in the blogosphere. In *KDD '09: Proceedings of the 15th ACM SIGKDD conference on Knowledge Discovery and Data mining*, 2009.

[12] K. Jarvelin and J. Kekalainen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.

[13] F. Radlinski, P. N. Bennett, B. Carterette, and T. Joachims. Redundancy, diversity, and interdependent document relevance. In *ACM SIGIR 2009 workshop*, 2009.

[14] F. Radlinski and S. Dumais. Improving personalized web search using result diversification. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval*, 2006.

[15] T. Sakai. New performance metrics based on multigrade relevance: Their application to question answering. In *NTCIR-4 Proceedings Open Submission Session*, 2004.

[16] M. Sanderson. Ambiguous queries: test collections need more sense. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval*, pages 499–506, 2008.

[17] K. Song, Y. Tian, W. Gao, and T. Huang. Diversifying the image retrieval results. In *Proceedings of the 14th annual ACM international conference on Multimedia*, pages 707–710, 2006.

[18] M. Strohmaier, M. Kroll, and C. Korner. Intentional query suggestion: making user goals more explicit during search. In *Proceedings of the 2009 workshop on Web Search Click Data*, pages 68–74, 2009.

[19] E. Vee, U. Srivastava, J. Shanmugasundaram, P. Bhat, and S. A. Yahia. Efficient computation of diverse query results. In *ICDE '08: Proceedings of the 2008 IEEE 24th International Conference on Data Engineering*, pages 228–236, 2008.

[20] R. W. White. Studying the use of popular destinations to enhance web search interaction. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval*, pages 159–166, 2007.

[21] Y. Yue and T. Joachims. Predicting diverse subsets using structural svms. In *ICML '08: Proceedings of the 25th International Conference on Machine Learning*, pages 1224–1231, 2008.

[22] C. Zhai, W. W. Cohen, and J. Lafferty. Beyong independent relevance: Methods and evaluation metrics for subtopic retrieval. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on research and development in information retrieval*, pages 10–17, 2003.

[23] C. Zhai and J. Lafferty. A risk minimization framework for information retrieval. *Information Processing and Management*, 42(1):31–55, 2006.

[24] B. Zhang, H. Li, Y. Liu, L. Ji, W. Xi, W. Fan, Z. Chen, and W.-Y. Ma. Improving web search results using affinity graph. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval*, pages 504–511, 2005.

[25] X. Zhu, A. Goldberg, J. V. Gael, and D. Andrzejewski. Improving diversity in ranking using absorbing random walks. In *NAACL-HLT '07: Proceedings of Human Language Technologies: the annual conference of the North American chapter of the Association for Computational Linguistics*, pages 97–104, 2007.

[26] C.-N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen. Improving recommendation lists through topic diversification. In *WWW '05: Proceedings of the 14th international conference on World Wide Web*, pages 22–32, 2005.