A Subtopic Taxonomy-Aware Framework for Diversity Evaluation

Fei Chen
Tsinghua University, P.R.China
chenfei27@gmail.com

Yiqun Liu
Tsinghua University, P.R.China
liuyiqun03@gmail.com

Min Zhang
Tsinghua University, P.R.China
z-m@tsinghua.edu.cn

Shaoping Ma
Tsinghua University, P.R.China
msp@tsinghua.edu.cn

Lei Chen
Tsinghua University, P.R.China
seavon@gmail.com

ABSTRACT

To evaluate search result diversification, which is supposed to meet different needs behind a same query, a number of evaluation frameworks are proposed and adopted by benchmarks such as TREC and NTCIR. These frameworks usually do not consider the subtopic taxonomy information. Many previous works on document ranking have shown that different kinds of information needs require different ranking strategies to be satisfied. It is thus necessary to involve subtopic taxonomy in the evaluation framework of search result diversification. In this paper, we propose a novel framework called the Subtopic Taxonomy-Aware (*STA*) framework to redefine the existing measures. Measures in this new framework take the subtopic taxonomy information into consideration for diversity evaluations. On the other hand, finding the optimal diversified results of many measures is proved as a NP-hard problem. We also propose a pruning algorithm which can decrease this problem to a computable search. Experiments based on both the TREC and NTCIR test collections show the effectiveness of our proposed framework.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: System and Software – *Performance evaluation*

General Terms

Algorithms, Measurement, Performance, Experimentation.

Keywords

subtopic taxonomy, diversified search, evaluation measures

1. INTRODUCTION

Evaluation measures have always been one of the most important and challenging topics in information retrieval research because of the parts they play in tuning and optimizing retrieval systems [13]. For the search result diversification task, many evaluation methods, such as *α-nDCG* [7], Intent Aware measures (e.g., *nDCG-IA*) [1], Expected Reciprocal Rank (e.g., *ERR* and *ERR-IA*) [9], and *D#-measures* [13, 15], have been proposed. Most measures use both document relevance, which traditional IR measures have considered, and the subtopic[1] coverage to evaluate diversified search. However, almost all these previous works treat different subtopics with the same methodology and ignore the fact that different strategies should be adopted to evaluate different kinds of information needs (or subtopics).

In the latest diversity evaluation studies, Sakai [12] has indicated that informational and navigational subtopics should be evaluated separately and differently to satisfy their different information needs. His work mainly focuses on proposing *DIN#-measures* and *P+Q#* to deal with this problem. However, we think works about how the subtopic taxonomy information would influence the diversity evaluation can be studied under a more general framework.

In this paper, supposing that (1) queries may have multiple subtopics; (2) for a given query, likelihood and taxonomy information of the subtopics underlying a query is available; and (3) graded relevance assessments between each subtopic-document pairs are available, we propose a Subtopic Taxonomy-Aware (*STA*) framework to redefine the existing measures for diversity evaluation in a general form. New evaluation measures under this framework are called *STA-measures*, which take not only the features of the corresponding existing measures but also the taxonomy information into account.

To the best of our knowledge, the *STA* framework is the first to try to generalize the use of subtopic taxonomy information in diversity evaluation. This means any taxonomy (e.g., the query intent or ODP-based content taxonomies) can be used to evaluate the diversified search result. The *STA* framework is also the first to introduce the decay function to redefine the document gain[2], according to both the subtopic taxonomy and different satisfactions among subtopics.

The contributions of this paper are:

1. A general framework called the *STA* framework, which incorporates the subtopic taxonomy information and can define new diversity evaluation measures.
2. A redefined document gain and several example decay functions, which are proposed based on the query topic taxonomy under the *STA* framework to help the original gain to be taxonomy-aware.
3. Thorough experiments and corresponding analyses, which are based on four completely different test collections using

[1] This paper refers to user intents underlying a query topic as subtopics.

[2] We define the gain value of a document [13] as "document gain" in this paper.

the Discriminative Power, Kendall's τ and τ_{ap} criteria to show the effectiveness of our framework.

4. A novel pruning algorithm, which can decrease the NP-hard optimal diversified result search to a computable time.

The remainder of this paper is organized as follows. In Section 2, we discuss related work and the existing diversity evaluation measures. Section 3 presents our new framework and its relationship to the existing measures. The proposed pruning algorithm is also described in this section. In Section 4, we provide the details of our four different test collections and the criteria used to evaluate new measures as experimental setup. In Section 5, we describe our experiments and provide corresponding analyses, by comparing the *STA-measures* with the existing measures. Finally, Section 6 presents our conclusions and directions for future work.

2. RELATED WORK

We first define symbols used in this paper. Search result evaluation requires assessing the document relevance with respect to a query topic. Documents are assessed with an integer level from 0 to h where 0 means irrelevant and h means the most relevant; h = 1 means a binary relevance assessment. Let d_r denote the document at rank r in the result list; we define $J(r) = 1$ if d_r is relevant at level x ($0 < x \leq h$), otherwise $J(r) = 0$. In light of this, the cumulative number of relevant documents is $C(r) = \sum_{i=1}^{r} J(i)$. Let $g(r)$ denote the document gain of d_r; $cg(r) = \sum_{i=1}^{r} g(i)$ thus means the cumulative gain at rank r. The gain and cumulative gain of the ideal ranked list are denoted as $g^*(r)$ and $cg^*(r)$, respectively.

The diversity evaluation requires assessments with respect to each subtopic instead of the topic, which differs from the traditional evaluations. Given a query q containing n subtopics, We denote the probability distribution of a subtopic i as $P(i|q)$, where $\sum_{i=1}^{n} P(i \mid q) = 1$. Document gains are assessed with respect to these n subtopics. We let $g_i(r)$ denote the gain of d_r with respect to subtopic i. Let $J_i(r)$ indicate whether d_r is relevant to subtopic i.

Clarke et al. [7] proposed a method called *α-nDCG* that accounts for the novelty and diversity of search results by explicitly counting the nuggets[1] in documents. Document gain in this measure depends on both the nuggets that the document contains and the nuggets covered by the already viewed documents. The Diversity Task of TREC 09 Web Track uses this measure as its main criterion.

Agrawal et al. [1] propose the Intent Aware (IA) framework to redefine the traditional measures. The concept of intent equals the subtopic notion used in this paper. These measures could better evaluate the diversified search result, as they explicitly consider the probability distribution of subtopics underlying a query.

Chapelle et al. [9] research the dependencies among documents in the result list. They assume that a user examines documents in the search result from top to bottom and, at each position, the user has a probability of being satisfied. When this happens, he stops and never examines any remaining document in the list. *ERR* is then defined as the reciprocal expectation of the probability that the user will stop at the current position. To use the *ERR* in diversity evaluation, Chapelle et al. also extended it to *ERR-IA* under the Intent Aware framework. The TREC 10 Web Track uses it as the main criterion in its Diversity task.

Sakai et al. [13, 15] propose an alternative framework called *D#-measures* to more intuitively evaluate the diversity of a search result list. The main idea is abandoning the separate calculation of measures for each subtopic, which is leveraged in the IA framework. Instead, the original document gains that are estimated in terms of each subtopic are linearly combined to define a new document gain (they called it *Global Gain*):

$$GG(r) = \sum_{i=1}^{n} P(i \mid q) g_i(r) \quad (1)$$

This definition is used to replace the document gains of the traditional measures to obtain new measures that are considered to evaluate the relevance and are referred to as *D-measures*. To evaluate the subtopic recall, Sakai et al. [13] also define the measure *I-rec*[2], which is the proportion of subtopics covered by the documents. With a document cutoff l, Sakai et al. [15] define the *D#-measures* by linearly combining the *D-measures* with *I-rec*:

$$D\#\text{-}measure@l = \lambda I\text{-}rec@l + (1-\lambda) D\text{-}measure@l \quad (2)$$

where λ is the tradeoff between the relevance and subtopic recall and is set to 0.5. The Document Ranking task of NTCIR 9 uses *D#-measures* as its official measures.

Sakai [12] suggests that diversity evaluation should distinguish between the navigational and informational subtopics. When the subtopic is navigational, the user wants to see only one particular web page, while the user is happy to see many relevant pages when the subtopic is informational. He indicates that the subtopic category should be considered and different measures should be leveraged to evaluate subtopics from different categories. Based on this, Sakai proposes *DIN#-nDCG* and *P+Q#*.

As described above, the existing measures mainly leverage the probability distribution of subtopics underlying a given query for diversity evaluation, and many measures, including *ERR-IA* and *D#-measures*, also consider the subtopic coverage of the already viewed documents. However, they fail to leverage the subtopic taxonomy information when evaluating. Although *DIN#-measures* and *P+Q#* utilize the query topic taxonomy (informational or navigational categories) in the evaluation process, they do not consider the influence of subtopic taxonomy in a general framework. In this paper, we abandon the subtopic taxonomy assumption, and propose the *STA* framework to redefine the traditional measures with any subtopic taxonomy.

3. SUBTOPIC TAXONOMY-AWARE EVALUATION FRAMEWORK

Benchmark-based studies such as TREC have shown that search tasks that focus on different kinds of query categories (e.g., homepage finding, topic distillation, named page finding) should utilize corresponding evaluation measures. For example, homepage finding tasks often adopt *MRR* while topic distillation tasks adopt *P@n*. Inspired by these findings, we believe that diversified search evaluation should also consider the taxonomy information of the subtopics underlying a given query. If we consider the most popular query topic taxonomy framework proposed by Broder et al. [2], which groups queries into navigational, informational and transactional categories, it is intuitive to conclude that one search target document is usually sufficient for a navigational subtopic. This is consistent with the evaluation of the Homepage Finding task in TREC [8], where the ranking of the first relevant document has vital importance. For informational or transactional sub-

[1] Clarke et al. [7] model the user's information needs as nuggets.

[2] Sakai et al. [13] rename *S-recall* [19] as *I-rec*.

topics, it is preferable to retrieve documents covering all detailed subtopic information.

We inherit all symbols defined in Section 2 for consistency. Moreover, if there are m subtopic categories under a certain taxonomy, we then define $P(k|i)$ as the probability that subtopic i belongs to each category k such that $\sum_{k=1}^{m} P(k \mid i) = 1$.

Here we consider that a subtopic may belong to one or more of the m categories with a probability distribution. It seems unreasonable because it is usually believed that each subtopic should only belong to a single category, since one subtopic stands for only one detailed user need. However, the reliability of this "one subtopic belongs to one and only one category" assumption is highly correlated to both subtopic mining granularity and subtopic annotation rules. We take topic 0015 in the Document Ranking subtask of the NTCIR 9 Intent Task as an example, whose query content is "Mozart". The subtopics given by NTCIR organizers contain: Mozart's music downloading, information about Mozart, opuses of Mozart, and schools named after Mozart, et al. Wiener Mozart Akademie is shown as a detailed example under the subtopic "schools named after Mozart". When users submit queries with this subtopic, some may be looking for the academy's homepage, but others may also be looking for information about this academy (history or something else). This subtopic (schools named after Mozart) may be either informational or navigational. People may argue that this subtopic can be divided into more detailed subtopics such that one is for finding the homepage, and the other is for information seeking. But this already explains that different granularities and rules of subtopic mining can cause one subtopic belongs to several categories. On the other hand, since the purpose of this paper is to construct a general diversity evaluation framework, we would like to define the problem with a general view such that it can be suitable for different kinds of subtopic taxonomies. For cases like TREC 09 or 10 (where a subtopic belongs to one and only one category), the probability distribution can be regarded as a specific function where $P(k|i) = 1$ if k is the only category, otherwise $P(k|i) = 0$ (For simplification, all the experiments of this paper are constructed in this way).

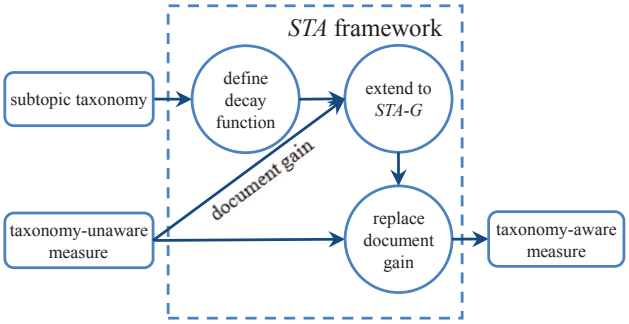

Figure 1. Function components of *STA* framework. Rectangles on the *left side* are the inputs and the rectangle on the *right side* is the output of the *STA* framework.

Figure 1 presents the main function components of the *STA* framework. The most important component in this framework is to extend the definition of document gain. It leverages the decay function to help itself characterize different subtopic taxonomies in a general form. The *STA* framework then defines new measures with the extended document gain such that these new measures can take the subtopic taxonomy into consideration.

3.1 The Subtopic Taxonomy-Aware Gain

To account for taxonomy information, we first extend the document gain definition, $g_i(r)$. Define S_r as the documents ranked higher than d_r in the result list. The extended document gain (called the subtopic taxonomy-aware gain, *STA-G*) is then defined as follows:

$$STA\text{-}G_i(r \mid S_r) = \sum_{k=1}^{m} P(k \mid i) g_i(r \mid k, S_r) \quad (3)$$

$g_i(r|k,s_r)$ is called the decayed gain: $g_i(r \mid k, S_r) = g_i(r) * f_k(i, S_r)$

where $g_i(r)$ is the document gain already defined in other diversity evaluation measures and $f_k(i,S_r)$ is a decay function used to discount $g_i(r)$ based on both the subtopic's category k and the subtopic coverage coming from documents in S_r. We discuss details about the decay function below. Equation (3) shows that *STA-G* is defined as a document gain calculated by linearly combining the decayed gain with $P(k|i)$.

3.2 Decay Function

In *STA* framework, any reasonable subtopic taxonomy can be used. The only thing we need to do is defining how the subtopic taxonomy and documents in S_r influence the current document gain. This influence is expressed by *decay function*. In a general form, decay function is a function of both the subtopic taxonomy and the influence coming from documents in S_r. As an example, we leverage (but is not limited to) the query topic taxonomy, which classifies queries into informational, navigational or transactional categories, as the subtopic taxonomy in this section and provide some *example* decay functions for each category.

Previous studies have proposed multiple decay functions for the traditional measures. Their properties have been studied primarily as functions of document ranking. For example, inspired by the discount function of *nDCG* proposed by Jarvelin et al. [13], Clarke et al. [7] suggest a logarithmic discount of $f(r) = 1 / \log_2(r+1)$. Chapelle et al. (2009) propose a linear discount of $f(r) = 1 / r$, and Clarke et al. [9] later suggest an exponential discount of $f(r) = \beta^{r-1}$, where $0 \le \beta \le 1$. Each form of these functions has its unique features. The logarithmic discount could obtain a slower decaying effect than the linear one. For the exponential discount, $f(r+1) / f(r) = \beta$, which is constant as r varies.

As examples, we only consider the number of relevant document as the influence coming from S_r. Define $C_i(S_r)$ as the number of documents relevant to subtopic i in S_r (i.e. $C_i(S_r) = \sum_{j=1}^{r-l} J_i(j)$). To discount the original document gain based on $C_i(S_r)$, we replace the document ranking r with $C_i(S_r)$ in the decay functions mentioned above:

$$\begin{aligned} f_{\log}(i,S_r) &= 1 / \log_2(C_i(S_r) + 2), \\ f_r(i,S_r) &= 1 / (C_i(S_r) + 2), \\ f_\beta(i,S_r) &= \beta^{Ci(S_r)} \end{aligned} \quad (4)$$

In the experiments conducted in this paper, β is assigned a value of 0.5 (i.e., the document that the assessor labels as relevant covers half of the remaining information need of a subtopic). Figure 2 shows how these three decay functions change along with the increase of $C_i(S_r)$.

As discussed in Section 3.1, $f_k(i,S_r)$ is also a function of the subtopic category. Recall that for the navigational subtopic, one relevant document is enough; but for the informational or transactional subtopics, several documents returned covering different facets are appropriate. Different decay strategies should thus be

leveraged for different categories. Compared with the informational or transactional subtopics, the navigational subtopic should be discounted much more after finding one relevant document. While for the informational or transactional subtopics, a more smooth decay with an increase in $C_i(S_r)$ should be applied.

Suppose we have a query with three subtopics denoted as s_1, s_2 and s_3, which are navigational, informational and transactional, respectively and have two result lists to be evaluated. The top three documents of both lists are d_1 (relevant to s_1), d_2 (relevant to s_2) and d_3 (relevant to s_3). The fourth document of the first list is relevant to s_1 while the one in the second is relevant to s_2 or s_3. In this scenario, the document gain of the former should be discounted much more than the document gain of the latter since the former cannot provide extra information and thus redundant. As a result, our measures prefer the second result list.

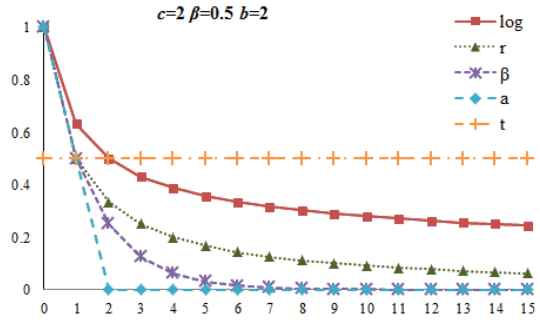

Figure 2. The curves of the example decay functions. The horizontal axis is $C_i(S_r)$, and the vertical axis is the decay function value. The labels *log*, *r*, *β*, *a* and *t* stand for the decay functions f_{log}, f_r, f_β, f_a and f_t, respectively.

Figure 2 shows that the decay functions in (4) present a smooth decay as $C_i(S_r)$ increases, which can be perfectly used as the decay function for the informational subtopic. As discussed above, however, the navigational subtopic can be well covered by the presence of one target page, making any other relevant pages redundant. We need a sharper decay function to characterize this. We introduce a new decay function here:

$$f_a(i,S_r)=\begin{cases}(c-C_i(S_r))/c, & 0\le C_i(S_r)\le c\\ 0, & else\end{cases} \quad (5)$$

where c is a parameter that stands for the number of relevant documents required by the navigational subtopic. In this paper, c is assigned a value of 2 (i.e., a maximum of two relevant documents is tolerable in the result list for a navigational subtopic). Figure 2 shows that it decays faster than the informational decay functions.

We also attempt to distinguish between the transactional and informational subtopics, although they are treated the same above. We think that users with transactional needs are always patient enough to find the pages that they think are the best within the top of the result list. These transactional pages can thus be equally treated and discounted, no matter how many documents are in the result list. This is reasonable because users always try their best to browse for the best sources within an acceptable range of pages before performing an effective download. In light of this, we define a decay function for transactional subtopics as follows: $f_t(i,S_r) = 1 / b$ where b is a constant assigned a value of 2 in this paper (i.e., the first relevant document of the transactional subtopic is half as important as the first relevance documents of subtopics in the other two categories).

Figure 2 shows the curves of each decay function with their parameters. This figure indicates that the decays of these functions increase from f_t, $f_{\log}$, f_r, and f_β to f_a. f_t could be considered a special case where the decay does not change. As Table 1 shows, f_t and f_a are taken as the decay functions of the transactional subtopic and the navigational subtopic respectively in the experiments. We discuss the effectiveness of the three functions, i.e., $f_{\log}$, f_r, and f_β, separately as the decay function of the informational subtopic in the experiment section. We must address again that these functions are *example* decay functions provided by ourselves based on the query topic taxonomy. Indeed, under the *STA* framework, one can define his own decay functions for any other subtopic taxonomy.

Table 1. Example decay functions for different subtopic categories.

Subtopic Category	Decay Function
informational	$f_{\log}$, f_r or f_β separately
navigational	f_a
transactional	f_t

3.3 Define Measures under *STA* Framework

Using the decay function definition, the *STA-G* in (3) can be rewritten as follows:

$$STA\text{-}G_i(r\mid S_r)=\sum_{k=1}^{m}P(k\mid i)g_i(r)f_k(i,S_r) \quad (6)$$

This document gain definition considers that a subtopic may belong to each category with some probability and leverages different decay functions to discount the original gain based on the subtopic's category. By replacing the document gain $g_i(r)$ of the *existing measures* with *STA-*$G_i(r|S_r)$, we obtain the corresponding *STA-measures*. The method proposed in this paper thus works as a framework to redefine measures such that the new measures can account for the subtopic taxonomy when evaluating the diversified search result. We call it *STA* framework. By considering a query has multiple subtopics, the *STA* framework captures the diversity of a query. By extending the document gain with decay functions defined according to the subtopic category, the *STA* framework accounts for the taxonomy information of subtopics. This framework guarantees that different decay strategies are leveraged for subtopics in different categories. Moreover, though the discussions about decay functions in this paper use the query topic taxonomy as the subtopic taxonomy, which classifies a subtopic into the informational, navigational or transactional category, we again address that our framework is not limited to this taxonomy. Equation (6) indicates that any taxonomy with corresponding decay functions defined can be applied to reformulating different *STA-Gs* and then to obtain different *STA-measures* in this framework.

We take the *nDCG-IA* as an example. After replacing its document gain with *STA-G*, we obtain the *STA-nDCG-IA*:

$$STA\text{-}nDCG\text{-}IA@l=\sum_{i=1}^{n}P(i\mid q)\frac{\sum_{r=1}^{l}STA\text{-}G_i(r\mid S_r)/\log(r+1)}{\sum_{r=1}^{l}STA\text{-}G_i^*(r\mid S_r)/\log(r+1)} \quad (7)$$

To convert *D#-measures* into the *STA* framework, we can first revise the *GG* defined in (1) as follows:

$$STA\text{-}GG(r,S_r)=\sum_{i=1}^{n}P(i\mid q)\sum_{k=1}^{m}P(k\mid i)g(r\mid i)f_k(i,S_r) \quad (8)$$

the *STA-D#-measures* are calculated using (2), with the replacement between the original *GG* and *STA-GG*.

Table 2 Test collection statistics.

	TRECDiv09	TRECDiv09+gr	TRECDiv10	NTCIR9Drk
Documents	ClueWeb09 (approx. one billion web pages) [4]			SogouT (approx. 138 million web pages in Chinese) [15]
Topics	50 topics (12 ambiguous; 38 faceted)		50 topics (27 ambiguous; 23 faceted)	100 topics
Subtopics	243 subtopics (177 informational; 66 navigational); mainly 199 subtopics with at least one relevant document.		218 subtopics (143 informational; 75 navigational); mainly 200 subtopics with at least one relevant document.	917 subtopics (811 informational; 62 navigational; 44 transactional); mainly 860 subtopics with at least one relevant document
Relevant	Across 50 topics: 4942; across 199 subtopics with relevant documents: 6499.		Across 50 topics: 6553; across 200 subtopics with relevant documents: 9006.	Across 100 topics: 12144; across 860 subtopics with relevant documents: 23571.
Submitted Runs	48 runs		32 runs	24 runs

If we let the decay function $f_k(i,S_r)$ in Equation (8) has the following specific form:

$$f_k(i,S_r)=\begin{cases} 1 & subtopic_i = informational \\ f_a(i,S_r) & subtopic_i = navigational \end{cases} \quad (9)$$

set parameter c in $f_a(i,S_r)$ (Equation (5)) to 1, and restrict that a subtopic should belong to one and only one category (i.e., $P(k|i)$ = 1 when k is the only category, otherwise $P(k|i)$ = 0), the *STA-D#-measures* in our framework degenerate to the *DIN#-measures* proposed by Sakai [12]. The *STA-D#-measures* can thus be regarded as a general form of this previous-proposed measure framework. We believe that the latter framework works well in most current diversity evaluation tasks because these tasks are based on the query topic taxonomy where queries are grouped into either navigational or informational category. However, the proposed *STA-D#-measures* are more flexible to subtopic taxonomies and they also provide choices of suitable decay functions to work with other specific taxonomies.

3.4 Pruning Algorithm

The decay function $f_k(i,S_r)$ can be defined to depend on the already viewed documents (e.g., decay functions of *STA-D#-measures*) or not (e.g. the simplified decay functions of *DIN#-measures*). When the function depends on the already viewed documents, finding the ideal list of *STA-measure* is an NP-Hard problem [7]. In most cases, a greedy algorithm is used to obtain an approximation:

Algorithm 1 Greedy Strategy

1 the ideal list $S^* \leftarrow \Phi$
2 while $|S^*| < L$ do
3 $d' \leftarrow \text{argmax}_{d \in R \setminus S}$ *STA-measure*$(S^* \cup \{d\})$
4 $R \leftarrow R \setminus \{d'\}$
5 $S^* \leftarrow S^* \cup \{d'\}$
6 end while
7 return S^*

In this paper, we propose a more effective approach to find the ideal list. The computational expensiveness of the exhaustive search is caused by the large candidate document set (D) or a big required list length (L). Practically, the candidate documents are usually several thousands and L is 1000 in most diversity tasks.

To decrease the complexity of the exhaustive search for the ideal list, we view this problem in the practical context of Information Retrieval. To reduce the number of candidate documents in a reasonable manner, we assume that there is no need to guarantee all the documents in the ideal list to be global optima at one time. Instead, we introduce a search window with fixed size on the L slots to make sure that the documents within the window are currently the optimal solution. Then the window moves until we cover all the documents in the L slots. For example, if we set the window size as 5. In the first step, the algorithm guarantees the documents from slots 1 to 5 are optimal. We then fix the documents from 1 to 5 and search for optimal documents from slots 6 to 10. Although the candidate document set D contains thousands of documents, every sub-list only comprises of 5 documents when the window size is 5. In each search step, we collect the next top 5 relevant documents for each subtopic in the unselected document set as the new candidate documents. If there are 9 sub-topics (which is NTCIR's maximum [14]) underlying a query, this method decreases the candidate document set D to 9*5 = 45 candidate documents. Considering that there may be common documents among sub-topics, the number of the new candidate documetns may be even smaller. If the window size is set to 1, the Search Window strategy equals to the greedy algorithm. Hence the smaller the window size is, the more effictive the search algorithm is but the more easily it obtains a local optima. As the top 10 documents are often cutted off to construct the evaluation, we set the window size to 5 as a tradeoff.

4. EXPERIMENTAL SETUP

4.1 Data Collection

In the experiment section, we compare the *STA-measures* with other measures on four different test collections to develop convincing conclusions. These test collections comprise runs submitted in different diversity tasks. These runs are created by different search systems based on two data sets: the ClueWeb data set (Category A) and the Chinese data set (SogouT[1]). Table 2 presents the statistics of these test collections.

4.1.1 TREC ClueWeb-Based Test Collections

In the Web Tracks of both TREC 2009 [4] and TREC 2010 [6], 50 different queries and their respective subtopics are provided for diversity evaluation. The subtopic taxonomy information is also provided, but only in two categories (informational and navigational). Subtopics labeled "inf" mean the user is seeking for information, while subtopics labeled "nav" mean the user is seeking for a previous-known specific page or site. The subtopic probability distributions are not given but are believed to exhibit equal probabilities in the TREC's diversity evaluation. Experiments based on these TREC test collections are thus conducted under the assumption that subtopics underlying a query are uniformly distributed. Moreover, TREC 09 and 10 annotate documents with a binary relevance, which means that each document is assessed as either relevant or irrelevant to a subtopic.

Base on the 50 queries provided by TREC 2009 and TREC 2010 Diversity tasks each, search systems of different competitors produce different runs. These runs constitute the TRECDiv09 and TRECDiv10 test collections, respectively. These two test collec-

[1] http://www.sogou.com/labs/dl/t-e.html

tions, whose queries and subtopics are mined in different ways and time intervals, are used to validate the effectiveness of experimental results across queries.

4.1.2 Graded Relevance of ClueWeb-Based Test Collection

To better compare *D#-measures* with other measures, Sakai et al. [15] re-annotated the TREC 09 Diversity Task dataset with tri-graded relevance, i.e. "relevant", "partially relevant" and "irrelevant". We construct experiments on this test collection (called TRECDiv09+gr) as complements to experiments done on the TREC ClueWeb Test Collection because graded relevance assessments can better demonstrate the advantages of graded relevance-aware measures such as *D#-nDCG*.

Runs used in this test collection are the same as the runs used in TRECDiv09. We also consider both uniform and non-uniform probability distributions of subtopics to determine their different influences. For the uniform distribution, all subtopics are equally likely, as in TRECDiv09. For the non-uniform distribution, we follow the work of Sakai et al. [13]: for a query with n subtopics, assume that the *j-th* subtopic has the probability $2^{n-j+1} / \sum_{k=1}^{n} 2^k$.

4.1.3 NTCIR 9 Test Collection

The test collections described in Sections 4.1 and 4.2 are based on ClueWeb dataset (Category A) adopted by TREC. To reach a more convincing conclusion, experiments should be conducted on different datasets. SogouT is a Chinese document collection used in the NTCIR 9 Document Ranking subtask of the Intent task. It is significantly different from the TREC dataset because it only includes Chinese Web pages, the size is smaller and the crawling was performed about half a year earlier than ClueWeb. Furthermore, this test collection contains 100 Chinese queries (topics), twice as many as TRECDiv09 or TRECDiv10. The probability distributions and user need descriptions are also provided for the subtopics of each query [16]. Documents of this collection are assessed with five-point scale graded relevance. However, the taxonomy information of the subtopic is not given. To utilize this dataset, we annotate each subtopic as informational, navigational or transactional categories based on their descriptions. Subtopics that describe the need for a certain page or site are labeled navigational, and those that describe the need for downloads or other transactions are labeled transactional. All other subtopics are labeled informational. The special subtopic, whose description is "Others" in the NTCIR 9 official assessed dataset, is labeled as informational. This special subtopic is used to accept user needs that the existing subtopics cannot cover. The NTCIR 9 official data collection is the only case among the four data collections that considers the "Others" subtopic. Experiments on this test collection can thus present a full range of possible subtopics hiding behind a query, which the others cannot do. We refer to this test collection as NTCIR9Drk.

4.2 The Evaluation of Diversity Evaluation Measures

Discriminative Power, Kendall's τ and τ_{ap} are the main criteria used to show the effectiveness of the newly proposed measures in the recent researches related to diversity evaluation [5, 12, 13, 15]. We also leverage them to evaluate the *STA-measures*.

4.2.1 Discriminative Power and Performance Difference Δ

Discriminative Power is a method confirmed and extended by Sakai et al. [11] to assess the discriminative power of various measures. By computing a significance test between every pair of submitted runs and counting the number of significantly different pairs, this method demonstrates how measures can be consistent across test collections (queries) and as a result, how often differences between systems can be detected with high confidence. For a significance test, we use the paired bootstrap method with B = 1000 bootstrap samples. The percentage of significantly different pairs among the B trials is called the Achieved Significance Level (ASL). The Discriminative Power also estimates the overall performance difference Δ required to achieve a given significance level α. Using the definition of the borderline between significance and non-significance as the *Bα-th largest* significance test value among the B trials, the performance Δ is defined as the largest borderline among all the run pairs. For the experiments in this paper, we fix the significance level at $\alpha = 0.05$ [11].

In each part of these experiments, we use the Discriminative Power at document cutoffs $l = 10$ and $l = 20$ as the main criteria to help us clearly compare the measures. This is because the top 10 or 20 documents in the result list are the most important in the diversity task and only the top 20 documents are pooled according to the TREC annotation rules [6].

4.2.2 The Kendall's τ and τ_{ap}

Kendall's τ is proposed and well-established to qualify the correlation between two rankings [10, 17]. Sakai et al. [12, 15] and Clarke et al. [5] also use it to measure the stability of experimental runs under different effectiveness measures. It takes values that range between [-1, +1], where +1 indicates perfect agreement and -1 the opposite. A τ value greater than 0.9 is thought to indicate that the rankings are nearly equivalent [3]. However, Kendall's τ does not suggest that items with high rankings should be considered more important than those with low rankings. To deal with this, Yilmaz et al. [18] proposed the τ_{ap} measure, which gives more punishment to errors at high rankings. Table 2 shows that the numbers of submitted runs in the four collections are different from each other. We randomly sample 20 runs for each collection as the different systems in the experiments. We thus take 20*19/2 = 190 different pairs to construct the statistical significance test when estimating the discriminative power.

4.3 The Subtopic Taxonomy and Decay Functions Used in Experiments

Although any reasonable taxonomy and decay function can be adopted in *STA* framework, we construct experiments using the query topic taxonomy and the example decay functions proposed based on this taxonomy to validate the effectiveness of *STA* framework. For simplification, we assume that a subtopic belongs to only one category in the experiments. This means the $P(k|i)$ values should only be 1 (if k is the annotated category) or 0 (otherwise). This is reasonable because subtopics should not be diversified in this informational-transactional-navigational subtopic category[1]. As Table 2 indicates, for the informational subtopic, we separately used f_{log}, f_r and f_β as the decay function. However, f_t and

[1] As the discussion in Section 3, whether a subtopic belongs to only one category is highly related to both the subtopic mining granularity and the subtopic annotations rules.

Table 3 Discriminative power (Column DP (%)) and performance Δ (Column Δ) of measures on test collections TRECDiv09, TRECDiv09+gr uniform, TRECDiv09+gr non-uniform, TRECDiv10 and NTCIR9Drk with α = 0.05.

		TRECDiv09				TRECDiv09+gr uniform				TRECDiv09+gr non-uniform				TRECDiv10				NTCIR9Drk			
		Cutoff 10		Cutoff 20		Cutoff 10		Cutoff 20		Cutoff 10		Cutoff 20		Cutoff 10		Cutoff 20		Cutoff 10		Cutoff 20	
	decay	DP	Δ	DP	Δ	DP	Δ	DP	Δ	DP	Δ	DP	Δ	DP	Δ	DP	Δ	DP	Δ	DP	Δ
D#-nDCG	none	70.00	0.16	72.11	0.2	69.47	0.14	72.63	0.16	68.95	0.16	73.16	0.14	67.37	0.17	68.42	0.16	64.21	0.09	61.58	0.09
	log	70.00	0.15	***74.21***	0.17	70.00	0.15	72.63	0.16	***72.11***	0.18	74.21	0.16	***68.42***	0.19	***69.47***	0.17	60.53	0.09	58.95	0.1
	β	70.00	0.15	***74.21***	0.19	69.47	0.16	72.63	0.16	71.05	0.15	***74.74***	0.17	68.42	0.17	68.95	0.19	62.63	0.1	57.89	0.09
	r	***70.53***	0.15	73.68	0.16	69.47	0.15	***73.16***	0.16	71.05	0.15	74.21	0.15	68.42	0.17	69.47	0.17	60.00	0.09	58.42	0.11
	s	***70.53***	0.15	71.58	0.16	***70.53***	0.14	72.63	0.14	70.00	0.15	72.63	0.15	66.84	0.15	66.32	0.16	***65.26***	0.09	***62.11***	0.09
D#-Q	none	70.53	0.17	72.63	0.14	70.53	0.14	72.11	0.15	70.53	0.14	***73.68***	0.14	65.79	0.2	***68.42***	0.16	***68.42***	0.1	***67.37***	0.08
	log	***72.11***	0.16	72.63	0.16	70.53	0.14	72.11	0.13	70.53	0.14	73.16	0.13	67.89	0.16	65.79	0.16	***68.42***	0.11	64.21	0.11
	β	71.05	0.18	72.63	0.15	***71.58***	0.15	***72.63***	0.14	70.53	0.15	71.58	0.15	66.84	0.17	66.84	0.16	65.26	0.1	64.21	0.12
	r	71.05	0.15	72.63	0.15	70.53	0.16	71.58	0.14	70.53	0.15	72.63	0.14	***68.42***	0.18	65.79	0.15	65.79	0.1	64.74	0.11
	s	70.53	0.15	72.11	0.18	70.00	0.16	71.58	0.17	70.53	0.14	72.11	0.14	67.89	0.17	65.79	0.15	66.84	0.1	65.79	0.1

f_a are used as the decay functions of the transactional subtopic and for the navigational subtopic respectively throughout the experiments. We can thus distinguish these experiments by the decay function used for informational subtopics and label them as "log", "r" and "*β*", respectively. Finally, we label the original *D#-measure* results as "none" (i.e., without any decay). Considering that *DIN#-measures* are the simplified cases of *STA-D#-measures* where the decay functions degenerate to indicators (Section 3.3), they can be thus taken as other forms of *STA-D#-measures* under the *STA* framework, whose decay functions differ from the *example decay functions*. We label the decay function of *DIN#-measure* as "*s*".

5. EXPERIMENT AND DISCUSSIONS

To show the effectiveness of the proposed *STA* framework, we selectively redefine the taxonomy-unaware measures: *D#-nDCG*, *D#-Q* [11] under the *STA* framework. This is because *D#-nDCG* and *D#-Q* stand for the state-of-the-art diversity evaluations and Sakai et al. [15] have already shown that these measures perform better than *ERR-IA*, *α-nDCG*, or other Intent Aware (IA) and *D#-measures*. Note that when the decay function is "*s*", *STA-D#-nDCG* degenerates to *DIN#-nDCG*, and *STA-D#-Q* degenerates to *DIN#-Q*.

Table 3 lists the experimental results on test collections TRECDiv09 (TRECDiv09, TRECDiv09+gr uniform and TRECDiv09+gr non-uniform), TRECDiv10 and NTCIR9Drk. These results contain the discriminative powers and overall performance difference Δs at the 0.05 significance level with document cutoffs 10 and 20, respectively. The "DP" Column in Table 3 presents the percentage of run pairs with $ASL > 0.05$. From this table, we can find:

1. The performances of the measures with "*β*", "log" and "r" example decay functions do not differ much among one another. As we fix the decay functions as f_a and f_t for the navigational and transactional subtopics, respectively and leverage the "*β*", "log" and "r" decay functions for the informational subtopic separately, the similar results imply that the capabilities of these decay functions for capturing the properties of user's informational needs are not so different. We can get similar conclusions by comparing results between the *STA-measures* using the example decay functions and the *DIN#-measures* (the "*s*" lines), which can be considered as other forms of *STA-measures* using the degenerate decay functions.

2. All measures on the graded-relevance-annotated data are as discriminative, as on the binary-annotated data. Similar conclusions can be reached between the uniform and non-uniform subtopic probability distributions.

3. The taxonomy-aware measures (including both *STA-measures* and *DIN#-measures*) are, in most cases, less discriminative on NTCIR9Drk but more discriminative on TREC-based test collections[1] than the taxonomy-unaware measures. This may be caused by the subtopic mining granularity and subtopic annotation rules. Note that the subtopics in TREC are first created by clustering user logs of a certain search engine, and then annotated with a clear subtopic category label. However, subtopics in NTCIR9Drk are annotated using pooling rules. Different participants mine and submit the subtopics, which are then pooled and annotated by the NTCIR official without considering their taxonomy. Some straightforward evidence is that the taxonomy-unaware measures (i.e., measures with the label "none") are poorly influenced when datasets are changed from TREC test collections to NTCIR9Drk.

Because the "log" decay function shows a stable effectiveness on discriminative power across datasets, we take "log" as the only decay function of the informational subtopic in the discussions below, though it is not the best on the TREC-based test collections. We provide the values of Kendall's τ and τ_{ap} for the original measures and measures with the "log" decay function in Table 4. The crucial information presented in Table 4 is that all these measures are highly correlated. Particularly, *STA-D#-nDCG* is stably correlated with the other measures.

Table 4 Kendall's τ and τ_{ap} on TREC10Div, decay=log, cutoff=10.

	D#-Q	*DIN#-nDCG*	*DIN#-Q*	*STA-D#-nDCG*	*STA-D#-Q*	*I-rec*
D#-nDCG	0.95/ 0.96	0.98/ 0.98	0.97/ 0.96	0.96/ 0.94	0.95/ 0.96	0.89/ 0.84
D#-Q	1/1	0.93/ 0.94	0.96/ 0.95	0.93/ 0.91	0.98/ 0.98	0.84/ 0.80
DIN#-nDCG	-	1/1	0.95/ 0.94	0.96/ 0.94	0.93/ 0.94	0.92/ 0.86
DIN#-Q	-	-	1/1	0.93/ 0.90	0.96/ 0.96	0.86/ 0.81
STA-D#-nDCG	-	-	-	1/1	0.93/ 0.91	0.92/ 0.88
STA-D#-Q	-	-	-	-	1/1	0.82/ 0.80

[1] There are some exceptions. For example, all decay functions used for *D-Q* at cutoffs 10 and 20, and all decay functions used for *D#-Q* at cutoff 20 perform less discriminatively than the "none" measures on TRECDiv10 test collections.

6. CONCLUSIONS AND FUTURE WORKS

In this paper, we propose a new framework, called *STA* framework, to generalize the taxonomy-aware diversity evaluation. This framework introduces decay function to define the *STA-G*, according to not only the subtopic taxonomy but also the influence of the already viewed documents. Under this new framework, any taxonomy with corresponding decay functions defined can be used to create *STA-measures* for diversified search evaluation.

Experiments leverage the topic taxonomy and the *example* decay functions proposed based on this taxonomy to validate the effectiveness of *STA* framework. These experiments are conducted on four different test collections, and show that *STA-measures* under the *STA* framework are comparable in terms of discriminative power with the *D#-measures* which represent state-of-the-art diversity evaluation measures.

For future work, we will propose a search result diversification method based on the *STA* framework. This method will incorporate both the subtopic taxonomical information and the *STA-G* of each document into the retrieval models.

7. ACKNOWLEDGMENTS

This work was supported by Natural Science Foundation (60903107, 61073071) and National High Technology Research and Development (863) Program (2011AA01A205) of China.

8. REFERENCES

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. 2009. Diversifying Search Results. In Proceedings of the Second ACM International Conference on Web Search and Data Mining. ACM, Barcelona, Spain (pp.5–14).

[2] A. Broder. 2002. A Taxonomy of Web Search. SIGIR Forum. 36(2), (pp.3–10).

[3] C. Buckley and E. M. Voorhees. 2004. Retrieval Evaluation with Incomplete Information. In Proceedings of ACM SIGIR 2004. ACM, UK (pp.25–32).

[4] C. L. Clarke, N. Craswell, and I. Soboroff. 2010. Overview of the TREC 2009 web track. In Proceedings of TREC 2009.

[5] C. L. Clarke, N. Craswell, I. Soboroff, and A. Ashkan. 2011. A Comparative Analysis of Cascade Measures for Novelty and Diversity. In Proceedings of ACM WSDM 2011. ACM, Hong Kong, China (pp.75–84).

[6] C. L. Clarke, N. Craswell, I. Soboroff, and G. V. Cormack. 2011. Overview of the TREC 2010 Web Track. In Proceedings of TREC 2010.

[7] C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Buttcher, and I. MacKinnon. 2008. Novelty and Diversity in Information Retrieval Evaluation. In Proceedings of ACM SIGIR 2008. ACM, Singapore (pp.659–666).

[8] N. Craswell, D. Hawking, R. Wilkinson, and M. Wu. 2004. Overview of the TREC 2003 Web Track. In Proceedings of TREC 2003 (pp.78–92).

[9] O. Chapelle, D. Metzler, Y. Zhang, and P. Grinspan. 2009. Expected Reciprocal Rank for Graded Relevance. In Proceedings of ACM CIKM 2009. ACM, Hong Kong, China. 621–630.

[10] M. Kendall. 1938. A New Measure of Rank Correlation. Biometrica. 30(1–2) (pp.81–89).

[11] T. Sakai. 2006. Evaluating Evaluation Metrics Based on the Bootstrap. In Proceedings of ACM SIGIR 2006. ACM, Seattle, WA, USA (pp.525–532).

[12] T. Sakai. 2012. Evaluation with Informational and Navigational Intents. In Proceedings of ACM WWW 2012. ACM, Lyon, France (pp.499-508).

[13] T. Sakai, N. Craswell, R. H. Song, S. Robertson, Z. Dou, and C.-Y. Lin. 2010. Simple Evaluation Metrics for Diversified Search Results. In 3rd International Workshop on Evaluating Information Access. Tokyo, Japan (pp.42–50).

[14] T. Sakai, Z. Dou, T. Yamamoto, Y. Liu, M. Zhang, and R. Song. Overview of the ntcir-10 intent-2 task. In Proceedings of NTCIR-10 Wordshop Meeting, 2013.

[15] T. Sakai and R. H. Song. 2011. Evaluating Diversified Search Results Using Per-intent Graded Relevance. In Proceedings of ACM SIGIR 2011. ACM, Beijing, China (pp.1043–1052).

[16] R. H. Song, M. Zhang, T. Sakai, M. P. Kato, Y. Q. Liu, M. Sugimoto, Q. L. W, and N. Orii. 2011. Overview of the NTCIR-9 INTENT Task. In Proceedings of NTCIR-9.

[17] E. M. Voorhees. 1998. Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness. In Proceedings of ACM SIGIR 98. ACM, Melbourne, Australia (pp.315–323).

[18] E. Yilmaz, J. A. Aslam, and S. Robertson. 2008. A New Rank Correlation Coefficient for Information Retrieval. In Proceedings of ACM SIGIR 2008. ACM, Singapore (pp.587–594).

[19] Cheng Zhai, William W. Cohen, John Lafferty. 2003. Beyond Independent Relevance: Methods and Evaluation Metrics for Subtopic Retrieval. In Proceedings of ACM SIGIR 2003. ACM, Toronto, Canada (pp.10–17).