Inf Retrieval (2013) 16:530–555
DOI 10.1007/s10791-012-9218-8

SEARCH INTENTS AND DIVERSIFICATION

Increasing evaluation sensitivity to diversity

Peter B. Golbus · Javed A. Aslam · Charles L. A. Clarke

Received: 22 May 2012 / Accepted: 18 December 2012 / Published online: 16 June 2013

Abstract Many queries have multiple interpretations; they are *ambiguous* or *underspecified*. This is especially true in the context of Web search. To account for this, much recent research has focused on creating systems that produce *diverse* ranked lists. In order to validate these systems, several new evaluation measures have been created to quantify diversity. Ideally, diversity evaluation measures would distinguish between systems by the amount of diversity in the ranked lists they produce. Unfortunately, diversity is also a function of the collection over which the system is run and a system's performance at ad-hoc retrieval. A ranked list built from a collection that does not cover multiple subtopics cannot be diversified; neither can a ranked list that contains no relevant documents. To ensure that we are assessing systems by their diversity, we develop (1) a family of evaluation measures that take into account the diversity of the collection and (2) a meta-evaluation measure that explicitly controls for performance. We demonstrate experimentally that our new measures can achieve substantial improvements in sensitivity to diversity without reducing discriminative power.

Keywords Diversity · Evaluation · Diversity difficulty

Portions of this material first appeared in the 2012 Diversity in Document Retrieval workshop (Golbus et al. 2012). This material is based upon work supported by the National Science Foundation under Grant No. IIS-1017903. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation (NSF).

P. B. Golbus (✉) · J. A. Aslam
College of Computer and Information Science, Northeastern University, 360 Huntington Ave., Boston, MA 02115, USA
e-mail: pgolbus@ccs.neu.edu

J. A. Aslam
e-mail: jaa@ccs.neu.edu

C. L. A. Clarke
School of Computer Science, University of Waterloo, 200 University Avenue West, Waterloo, ON N2L 3G1, Canada
e-mail: claclark@plg.uwaterloo.ca

1 Introduction

Information retrieval research traditionally assumes that a given query can be associated with a single underlying user intent, or information need. In reality—especially in the context of Web search—users with very different intents may enter identical queries. In many cases, queries may be *ambiguous*, with multiple unrelated interpretations (Clarke et al. 2009). For example, a user entering the query "zeppelin" may be interested in either the band or the type of airship. Even when a query is unambiguous it may be *underspecified*, and may not precisely express the user's information need (Clarke et al. 2009). For example, a user entering the query "led zepplin" may be seeking a discography, biographies of the members, and/or news about a possible reunion. When a query gives rise to many possible interpretations, the ideal ranked result list should be both *diverse*—it should cover as many search intents as possible—and *novel*—each new document should provide previously unseen information (Clarke et al. 2008). Several recent research efforts have proposed algorithms that support ambiguous and underspecified queries by diversifying search results (Santos et al. 2011; Vargas et al. 2012; Zuccon et al. 2012).

Evaluating these algorithms requires effectiveness measures that appropriately reward diversity in the result list. Many such measures have been proposed, necessitating the need for meta-evaluation analyses. A key component frequently used in such meta-evaluations is *discriminative power* (Sakai 2006), which assesses how sensitive measures are to changes in ranked lists. While discriminative power provides a valuable tool, it was originally designed for ad-hoc performance measures. Discriminative power alone does not tell us if the sensitivity is due to diversity, or if it is due to other factors.

Diversity necessarily depends on the collection over which the search is being run. If we search back-issues of the Journal of Aerospace Engineering for the query ”zeppelin,” we are unlikely to find many references to the band. When a collection only covers a single interpretation, even the best search engine will be unable to create a diverse ranked list. Alternatively, consider searching a corpus like Wikipedia for "led zeppelin." In this case, any relevant document is likely to be at least partially relevant to several intents, and almost any system will produce a diverse ranked list. Diversity also depends on a system's performance at basic ad-hoc retrieval—how many documents are relevant to any reasonable intent, especially at the top of the ranked list. Poor ad-hoc performance implies poor diversity; a system that returns few documents relevant to any intent cannot present a diverse ranked list to the user. To isolate diversity from these confounding factors, we define (1) a new family of diversity evaluation measures that explicitly account for the diversity of the collection and (2) a meta-evaluation measure that can assess the sensitivity of diversity evaluation measures when applied to artificial ranked lists with perfect performance.

In order to isolate a system's diversity from that of the collection, we develop a notion analogous to query difficulty that measures the diversity present in a collection, independent of any ranked list. *Diversity difficulty* (Sect. 4) measures this property at the topic (i.e., query) level, and *subtopic miss rate* (Sect. 5) measures this property at the subtopic (i.e., interpretation) level. In Sect. 6, we show how to incorporate the diversity of the collection into a new family of diversity evaluation measures, α#-IA measures. We demonstrate experimentally that these measures prefer different systems than existing measures, and have slightly higher discriminative power.

In order to isolate a measure's sensitivity to diversity from its ad-hoc performance, in Sect. 7 we consider artificial ranked lists created by randomly permuting relevant documents. We show experimentally that, as measured by discriminative power, no measure is sensitive to the changes in these lists. Therefore, we develop a new meta-evaluation measure, *document selection sensitivity*, which is able to distinguish between evaluation measures applied to perfect-performance ranked lists.

2 Related work

Traditionally, information retrieval followed the Probabilistic Ranking Principle (Robertson 1977), which dictates that documents should be ranked by their probability of relevance to the user's intent, with the simplifying assumption of independent document relevance. Zhai et al. (2003) explicitly rejected the assumption of independent relevance to define the *subtopic retrieval* problem, which came to be known as diversity. Most existing diversification systems are inspired (Santos et al. 2011) by Carbonell and Goldstein's Maximal Marginal Relevance method (Carbonell and Goldstein 1998), which contrary to the PRP, iteratively selects documents that are most similar to the query and least similar to the documents previously shown to the user. Indeed, Chen and Karger (2006) showed that the PRP is actually sub-optimal if the user is interested in a limited number of relevant documents, rather than all relevant documents.

Zhai et al. also invented several evaluation measures that can be used to measure diversity. Subtopic recall, the percentage of subtopics covered by a ranked list, is still in prominent use, usually referred to as S-Recall or I-Rec. Recently, research in ad-hoc retrieval evaluation has been exploring measures that are driven by user models. Craswell et al. (2008) introduced the *cascade model* of user behavior, under which users are assumed to scan ranked lists from the top down, stopping when their need is satisfied or their patience is exhausted. This model has been incorporated into evaluation measures such as: expected reciprocal rank (ERR) (Chapelle et al. 2009), rank-biased precision (RBP) (Moffat and Zobel 2008), and expected browsing utility (EBU) (Yilmaz et al. 2010). Chapelle et al. (2009) demonstrated experimentally that their cascade model is better correlated with user behavioral metrics than traditional measures. Clarke et al. (Clarke et al. 2008) modified nDCG (Järvelin and Kekäläinen 2002) by using the diminishing returns of cascade measures to model the user's tolerance for redundancy in a ranked list. By penalizing redundancy, α-nDCG rewards both novelty and diversity.

As originally defined, Clarke et al.'s cascade measures treated all documents and subtopics equally. Agrawal et al. (2009) proposed the IA measure family that computes the weighted average of ad-hoc metrics computed separately for each intent. These measures incorporate both degrees of relevance, known as graded relevance, and the likelihood that a user is interested in a particular subtopic, known as intent probability. As an example of intent probability, suppose that more users who enter the query "zeppelin" are interested in the band than the mode of travel. Then systems should receive higher rewards for retrieving documents that refer to the former than the later. Cascade measures were extended to also utilize graded relevance and intent probabilities, as well as to use other discount functions (Clarke et al. 2009, 2011).

A drawback of IA measures is that they tend to prefer systems that perform well on the most likely subtopics over systems that are more diverse (Clarke et al. 2011; Sakai et al. 2010). Partially in an attempt to correct this problem, Sakai et al. (2010) introduced the D# family of measures. These measures compute the weighted average of each document's gain for each subtopic, which the authors call "global gain." This formulation can be derived from the probabilistic ranking principle (see Sect. 6), and has the added advantage that the greedily constructed ideal rank list is optimal, unlike for α-nDCG where it is an approximation. The authors refer to measures that use global gain as D measures. D# measures are D measures linearly combined with S-Recall, and are more highly correlated with subtopic coverage than IA measures. Noting that D# measures seemed to perform differently on subtopics depending on their taxonomy, i.e. *navigational* versus *informational* (Broder 2002), Sakai (2012) developed additional measures in the style of D# that explicitly take subtopic taxonomy into account. One of these, the P+Q# measures, uses what we would call a #-IA measure in the sense of Sect. 6.2 when the intent is informational.

As we show in Sect. 7, existing measures are dominated by ad-hoc performance—systems are ranked primarily by their ability to retrieve relevant documents and only secondarily by their ability to diversify them. Such a result is consistent with other work in the literature. Consider, for example, an experiment from recent work by Santos et al. (2012). In that work the authors investigate the relative impact of increasing subtopic coverage versus reducing redundancy. In one particular experiment, they show that taking a quality ad-hoc run and diversifying it using state-of-the-art algorithms can increase the α-nDCG@100 from roughly 0.35 to 0.45. However, removing all non-relevant documents from the list without any attempts at diversification increases the α-nDCG@100 to almost 0.6. This implies that the best way to maximize α-nDCG may very well be to continue to improve ad-hoc retrieval.

In light of this, how do we analyze evaluation measures? The two main approaches to meta-evaluation used in the literature are “discriminative power,” based on hypothesis testing, and ”intuitiveness,” which analyzes the ideal ranked lists produced by optimizing for each measure. Discriminative power, which uses a Bootstrapped *t* test, was first introduced by Sakai et al. (2006) as a meta-evaluation measure for ad-hoc performance metrics. It has since been used extensively to evaluate diversity measures (Clarke et al. 2011; Sakai 2012; Sakai et al. 2010; Sakai and Song 2011). Intuitiveness (Sakai 2012; Sakai et al. 2010; Sakai and Song 2011) attempts to view evaluation measures through the eyes of a user.

In this work, we adopt the opposite approach, viewing measures from the perspective of the collection. It is our hypothesis that by leveraging as much information as we can, no matter how opaque to the end-user, we will be best able to distinguish between systems. While a user neither knows nor cares whether a particular query is hard, we believe that if we can find systems that are still able to perform reasonably on the most difficult queries, they will tend to best satisfy users over all. It remains to be seen the extent to which our collection-oriented perspective agrees with Sakai et al.’s user-oriented perspective, or with actual user preference.

Another avenue we explore in this work is the difficulty of diversity queries. A great deal of work has already been done to distinguish hard queries for ad-hoc performance. We give a very brief overview of the work in the area of query difficulty prediction. We direct the interested reader to Carmel and Yom-Tov (2010), in whose taxonomy we situate our work. At the highest level, this schema first divides query difficulty prediction by whether the analysis is performed pre- or post-retrieval.[1] Pre-retrieval prediction is further divided by whether the analysis is statistical (Hauff 2010) or linguistic (Mothe and Tanguy 2005) in nature. Post-retrieval prediction is split into three categories: clarity, score analysis, and robustness. Clarity analyzes the difference between the top-retrieved documents and the collection as a whole (Carmel et al. 2006; Cronen-Townsend et al. 2002, 2006), and the score analysis measures how much the document scores reported by a system vary at the

[1] These categories are due to He and Ounis (2004)

top of the list and across the corpus as a whole (Shtok et al. 2009; Zhou and Croft 2007). Robustness measures the extent to which retrieval is affected by perturbations. These perturbations can be to the query (Vinay et al. 2006; Yom-Tov et al. 2005; Zhou and Croft 2007), to the document set (Vinay et al. 2006; Zhou and Croft 2006) or to the retrieval systems (Aslam and Pavlu 2007). Our approach to measuring diversity difficulty at the topic and subtopic level is to consider the output of systems that pick relevant documents at random. Even though we do not use actual IR systems, this falls in the category of post-retrieval robustness as measured by perturbing systems.

3 The TREC diversity task

Since 2009, the TREC Web Track (Clarke et al. 2009, 2010, 2011) has included a diversity task alongside the traditional ad-hoc task. The track organizers constructed 50 topics for each collection. For each topic, the organizers created a number of subtopics corresponding to example search intents by extracting information from the logs of a commercial search engine. Each subtopic is given a type, “navigational” or “informational,” denoting whether the user is interested in finding a specific web page or any web page with the correct content. Figure 1 presents two topics from the 2011 collection and their subtopics. Topic 114 is “faceted,” i.e. ”underspecified” in the same sense as our “led zeppelin” example; we know what an adobe indian house is, but we do not know which facet of this broad topic the user is interested in. Topic 140 is “ambiguous,” similar to our “zeppelin” example. There are many high schools named East Ridge; without additional information, there is no way of knowing which one the user intended.

Track participants were given the queries (and not the subtopics) associated with each topic. Systems were run on the ClueWeb09 corpus,[2] a general web crawl from 2009 containing approximately 1 billion documents. The submissions were pooled and judged to a depth of 20 (2009, 2010) or 25 (2011). Hired assessors made binary relevance judgments with respect to each subtopic.[3] All of our evaluations will be performed using these relevance assessments.

In this paper, we focus on the 2010 and 2011 collections, since participants in those years had time to work with the 2009 data to better understand how to diversify runs. Where possible, we considered the two as a single collection with 100 topics.

4 Diversity difficulty

The amount of diversity present in a ranked list is affected by the amount of diversity present in the collection. Therefore, when evaluating systems for diversity, it is necessary to control for the diversity of the collection. For a specific topic and corpus, query difficulty is a measure of how well a reasonable search engine can be expected to perform ad-hoc retrieval. In this section, we introduce an analogous notion for diversity, **diversity difficulty**, which assesses the amount of diversity present in the collection. Like query difficulty, diversity difficulty is defined with respect to a topic and a corpus, and independent of

[2] lemurproject.org/clueweb09/.

[3] In 2011, the judgments were actually graded. For this work, we consider any document with relevance grade greater than zero to be relevant and any document with a relevance grade of zero or marked as spam to be non-relevant.

```
<topic number="114" type="faceted">
  <query>adobe indian houses</query>
  <description>
    How does one build an adobe house?
  </description>
  <subtopic number="1" type="inf">
    How does one build an adobe house?
  </subtopic>
  <subtopic number="2" type="inf">
    information about Indian tribes that used adobe houses
  </subtopic>
  <subtopic number="3" type="nav">
    I'd like to order books or videos/CDs about how to construct adobe buildings.
  </subtopic>
</topic>

<topic number="140" type="ambiguous">
  <query>east ridge high school</query>
  <description>
    demographics of East Ridge High School in Lick Creek, Kentucky
  </description>
  <subtopic number="1" type="inf">
    demographics of East Ridge High School in Lick Creek, Kentucky
  </subtopic>
  <subtopic number="2" type="nav">
    home page for East Ridge High School in Chattanooga, Tennessee
  </subtopic>
  <subtopic number="3" type="inf">
    information about the sports program at East Ridge High School in Clermont, Florida
  </subtopic>
  <subtopic number="4" type="inf">
    description of the sports facilities at East Ridge High School in Woodbury, MN
  </subtopic>
</topic>
```

Fig. 1 Examples of TREC 2011 web track diversity task topics

any ranked list. Diversity difficulty is defined in Sect. 4.1. The diversity difficulties of the TREC 2010 and 2011 corpora are analyzed in Sect. 4.2.

4.1 Definition

Imagine a collection and a topic with ten subtopics and 1,009 relevant documents. One of these subtopics, subtopic A, is covered by 1,000 different documents. Subtopics B through J are each covered by only one relevant document. It is possible to generate a diverse ranked list that covers all ten subtopics, but it is difficult. A system would need to order those nine documents relevant to subtopics B–J high in the list—the equivalent of finding a handful of "needles" in the "haystack" of 1,000 documents relevant to subtopic A. However, imagine a different collection and the same topic. In this collection, there are large numbers of documents relevant to each subtopic, or perhaps there are large numbers of documents relevant to multiple subtopics. In this collection, it would be easy to produce a diverse list. In fact, almost any list with good performance should exhibit diversity. However, this topic exhibits the same maximum amount of diversity in both collections—each of the 10 subtopics can be covered by some ranked list of ten documents. One could also imagine a third collection where subtopics A through I are each covered by many documents, each of which cover many subtopics, but there are no documents whatsoever relevant to subtopic J. In this collection, it would still be easy to create a diverse ranked list, but the maximum diversity is smaller than in the first two collections. One might argue that this simply means that subtopic J should be disregarded, and that this third collection is just as diverse. We will argue in Sect. 4.2 that this depends on how the collection was created, and the purpose it is intended to serve.

We consider diversity difficulty to be a function of the two factors previously discussed: (1) the maximum amount of diversity achievable by any ranked list, and (2) the ease with which a system can produce a diverse ranked list.[4] When the maximum amount of diversity achievable by any system is small, the topic has little diversity. When the maximum amount of diversity is large but it is hard to create a diverse list, the topic is somewhat more diverse. Finally, if the maximum amount of diversity is large and a system created at random will come close to achieving it, the topic is diverse.

Given a topic, S-Recall@k (Zhai et al. 2003) is the percentage of subtopics covered by a list at rank k. The S-recall of a set of documents is the same for any ranked list of those documents. We consider the maximum amount of diversity (denoted d_{max}) for a topic to be the **Maximum S-Recall** for any set of documents in the corpus. Let ξ represent the minimal cutoff k at which d_{max} can be achieved, i.e. the minimum number of documents that cover the maximum number of subtopics. Unfortunately, computing ξ can be shown to be NP-hard by reduction from the set covering problem (Carterette 2009). In this work, we use a greedy approximation.

Once we know ξ, imagine the random experiment of selecting ξ relevant documents from the corpus and measuring the S-Recall. The expectation of this experiment is analogous to the S-Recall of a system that performs ad-hoc retrieval perfectly, yet does not attempt to diversify its results. We use the **Expected S-Recall@$\boldsymbol{\xi}$** (also denoted d_{mean}) to measure how easy it is to create a diverse list. Let M be the number of subtopics, R_i be the number of documents relevant to subtopic i, and R_T be the number of documents relevant to at least one subtopic. Then d_{mean} can be approximated as

$$d_{mean} \approx 1 - \frac{\sum_{i=1}^{M}\left(1 - \frac{R_i}{R_T}\right)^{\xi}}{M}. \tag{1}$$

We note that while d_{mean} can be computed directly, we use an approximation that actually models documents sampled with replacement. Therefore this approximation can be poor when there are few relevant documents for a subtopic, e.g. when the subtopic is navigational. We define diversity difficulty, dd, as the harmonic mean of d_{max} and d_{mean},

$$dd = \frac{2d_{max}d_{mean}}{d_{max} + d_{mean}}. \tag{2}$$

Since S-recall is a percentage of subtopics, diversity difficulty ranges between zero and one. It is large for diverse queries where there are many subtopics and an arbitrary ranked list is likely to cover many of them. It is small for queries lacking in diversity where there are either few subtopics, or there are many subtopics but they are unlikely to be covered.

[4] Note that our definition of diversity difficulty will actually be a description of diversity ”easiness” in that larger values indicate topics on which systems should do well. This is similar to e.g. query average average precision (Aslam and Pavlu 2007). We choose to call this diversity difficulty rather than diversity easiness to emphasize the similarity to query difficulty prediction.

Table 1 Examples of subtopic coverage and diversity difficulty in TREC 2010 and 2011 topics

Topic ID	Title	# Rel Docs	Subtopic						*dd*
			1	2	3	4	5	6	
143	arkadelphia health club	25	25	21	–	–	–	–	0.994
86	bart sf	82	78	62	60	–	–	–	0.977
125	butter and margarine	132	110	47	13	–	–	–	0.735
73	neil young	156	69	52	28	19	–	–	0.730
60	bellevue	313	254	47	16	11	4	4	0.481
57	ct jobs	261	261	14	5	2	–	–	0.449

4.2 TREC collections

A commercial web search engine—which, in theory, indexes the entire web—must retrieve relevant documents for every search intent, no matter how rare. In this context, it is important to find those intents that the search engine is unable to satisfy so that the situation can be rectified. TREC collections, however, are more artificial. Designed to evaluate search engines, they consist of a first tier web crawl and topics created by visually inspecting the search logs of a commercial search engine. In this context, there are often uncovered subtopics with no relevant documents. These subtopics may not have represented common user intents, or documents pertaining to them may be missing from the crawl. Therefore, *for TREC collections only*, we restrict our attention to subtopics that are actually covered by relevant documents. However, this changes the meaning of diversity difficulty. Due to the collection we are using, *in these experiments*, the Maximum S-Recall will be 1 for any topic. In this case, topics will be considered diverse, *i.e. dd* is large, if and only if an arbitrary ranked list is likely to cover all subtopics, independent of the number of subtopics.

Measuring the diversity difficulty of the TREC 2010 and 2011 topics, we see that *dd* does in fact measure the diversity of topics. Table 1 shows several topics and the number of relevant documents for each subtopic. Topics 143 (“arkadelphia health club”) and 86 (“bart sf”), have a non-negligible and roughly equal number of relevant documents for each subtopic. These are very diverse topics, and they have very high diversity difficulty scores of almost 1. Topics 125 (”butter and margarine”) and 73 (“neil young”) each cover all subtopics with many documents, but some subtopics are covered by many more documents than others. They have some diversity, which is reflected in their diversity difficulty scores of about 0.75. Topics 60 (“bellevue”) and 57 (“ct jobs”) both have dominant subtopics that are far more covered than the others, as well as subtopics that are barely covered. These topics have little diversity, and low diversity difficulty scores that are less than 0.5.

Figure 2 shows a histogram of the diversity difficulty of the topics in the combined TREC 2010 and 2011 collection. Table 2 shows the minimum, maximum, and mean diversity difficulty values for each year. Using diversity difficulty, we can see that the TREC 2010 and 2011 collections were diverse, with 2011 being somewhat more so.

5 Subtopic miss rate

Because it is necessary to average system evaluations over topics to control for natural variations within a collection, diversity difficulty tells us which topics are naturally more

Fig. 2 Histogram of diversity difficulties of the topics in the combined TREC 2010 and 2011 collection. The larger the value, the easier it is to create a diverse ranked list for that topic

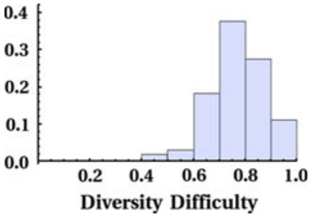


Table 2 TREC 2010 and 2011 collection diversity difficulty statistics

	Min	Max	Mean
2010	0.449	0.994	0.727
2011	0.643	0.977	0.809

diverse than others. However, for individual topics, there is variation among subtopics as well. In this section, we present **subtopic miss rate**, which, for each topic, measures the relative prevalence of documents relevant to each subtopic.

For a given topic, consider drawing relevant documents at random. Subtopics containing large numbers of relevant documents will be covered early and easily—these are the ”easy” subtopics, likely to be covered by any system with reasonable ad-hoc performance. However, subtopics with few relevant documents are “harder” and will likely be covered early and well by only high-quality diversity systems.

We define the **subtopic miss rate** of subtopic *i* at rank *k* as the probability of drawing *k* relevant documents at random and failing to cover that subtopic, normalized with respect to all of the subtopics for that topic. This forms a distribution for each topic, with probabilities corresponding to each subtopic's relative difficulty.

It is not strictly necessary to normalize these probabilities into a distribution. We do so to emphasize the relative importance of each subtopic to evaluation, rather than any universal notion of difficulty. The drawback to this approach is that the normalized probabilities are not comparable across topics.

Let M be the number of subtopics, R_i represent the number of documents relevant to subtopic i, and R_T represent the total number of documents relevant to at least one subtopic. The subtopic miss rate, *smr*, of subtopic i at rank k can be approximated as

$$smr_i^k \approx \frac{\left(1 - \frac{R_i}{R_T}\right)^k}{\sum_{j=1}^{M}\left(1 - \frac{R_j}{R_T}\right)^k}. \tag{3}$$

While smr_i^k can be computed directly, to simplify computation, we approximate, asserting that documents are sampled with replacement. Again, this is a poor assumption for subtopics with a small number of relevant documents. If no rank is specified, we define the *smr* of a subtopic as the *smr* at rank ξ,

$$smr_i = smr_i^{\xi}, \tag{4}$$

where ξ is the minimum rank at which all subtopics can be covered.

If a subtopic is covered by all relevant documents, then every reasonable system should be able to cover it. The miss rate of this subtopic is zero at any rank; the subtopic is of no interest and should be ignored. However, if a subtopic is covered by only a relatively small number of relevant documents while most other subtopics are covered heavily, then the miss rate could approach one. This implies that this subtopic will be very useful in differentiating between the best systems.

Table 3 shows the subtopic miss rate of several TREC topics. Topic 60 has a small diversity difficulty score, meaning that it is a topic with little inherent diversity. Subtopic one has a very small subtopic miss rate—it is very unlikely to be missed by any ranked list. The remaining five subtopics all have very similar subtopic miss rates. They are all equally unlikely to be covered. This implies that the first subtopic is not useful for evaluation, and that systems should be measured by how well they cover the other five subtopics. Topic 73 has an intermediate diversity difficulty score, and is therefore a topic that is neither particularly diverse nor lacking in diversity. The subtopic miss rates are less similar, showing that several subtopics are highly likely to be covered and several subtopics are less likely to be covered. Each subtopic should contribute differently to evaluation. Topic 86 has a high diversity difficulty score; it is a diverse topic. This topic has a single "dominant" subtopic with a very high miss rate. All systems should be expected to satisfy the more common user intents, and therefore they should be evaluated largely by whether or not they are able to satisfy this rare one.

6 Diversity evaluation measures

We have shown how to quantify the amount of diversity present in a collection at the levels of topic and subtopic. In Sect. 6.1, we briefly give definitions for several existing evaluation measures. In Sect. 6.2, we introduce a new family of evaluation measures, the α#-IA measures. In Sect. 6.3, we show how our new measures can be made to explicitly account

Table 3 Example subtopic miss rates of TREC topics

Topic ID	Title	*dd*	ξ	Subtopic	Rank			
					ξ	5	10	20
60	"bellevue"	0.481	3	1	0.002	0.000	0.000	0.000
				2	0.143	0.113	0.061	0.016
				3	0.199	0.196	0.182	0.144
				4	0.209	0.213	0.215	0.202
				5	0.224	0.239	0.271	0.319
				6	0.224	0.239	0.271	0.319
73	"neil young"	0.730	2	1	0.141	0.050	0.007	0.000
				2	0.202	0.122	0.040	0.003
				3	0.306	0.344	0.321	0.204
				4	0.351	0.484	0.632	0.793
86	"bart sf"	0.977	1	1	0.087	0.00	0.000	0.000
				2	0.435	0.383	0.278	0.129
				3	0.478	0.617	0.722	0.871

All subtopics tend to have similar rates for non-diverse topics (small diversity difficulty scores), whereas diverse topics (high diversity difficulty scores) tend to have "dominant" subtopics. Recall that ξ is the minimum rank at which all coverable subtopics can be covered by any ranked list for collection diversity at the topic and subtopic levels. Finally, we demonstrate experimentally that these measures prefer different systems than existing measures (Sect. 6.4), and that they have slightly higher discriminatory power (Sect. 6.5)

6.1 Intent-aware, cascade, and D#-measures

Following Zhang et al. (2010), we note that most measures can easily be described as the cross-product of a gain vector with a discount vector, normalized in some fashion. Following Clarke et al. (2011), we begin by highlighting three particular functions for producing discount vectors based on ad-hoc performance measures.

1. **S-Recall** (Zhai et al. 2003): S-Recall does not fit into this framework. Let X be the number of subtopics covered by at least one document at or before rank k. Assume a topic has M subtopics. Then

$$\text{S-Recall@}k = \frac{X}{M}. \tag{5}$$

2. **Cascade** (Clarke et al. 2008, 2009, 2011): These methods use a cascading gain function where a document's gain with respect to some subtopic is decreased each time the subtopic is encountered. To define the gain function, let I_i^r be an indicator variable representing whether the document at rank r is relevant to subtopic i. Let

$$c_i^k = \sum_{r=1}^{k-1} I_i^r \tag{6}$$

represent the number of documents relevant to subtopic i seen prior to rank k. Let g_i^k be the relevance grade, or a function thereof, of the document at rank k with respect to subtopic i. Let w_i represent the probability that a user's intent is subtopic i. Then the gain of the document at rank k is

$$\text{Gain}(k) = \sum_{i=1}^{M} w_i \times g_i^k (1-\alpha)^{c_i^k}, \tag{7}$$

where α is a parameter in [0, 1] which models the users tolerance for redundancy. The gain can be combined with any discount vector from Table 4. Normalization is performed relative to a single ideal ranked list. A drawback of the Cascade measures is that the ideal ranked list required for normalization is NP-hard to compute, and is therefore usually approximated.

3. **Intent aware** (Agrawal et al. 2009): Intent aware measures model diversity by computing the weighted average of an evaluation measure with respect to each subtopic. As an example of an intent aware measure, consider nDCG-IA. Let nDCG_i represent nDCG (Järvelin and Kekäläinen 2002) evaluated with respect to subtopic i. Then the intent aware measure nDCG-IA would be:

$$\text{nDCG-IA@}k = \sum_{i=1}^{M} w_i \times \text{nDCG}_i\text{@}k. \tag{8}$$

Notice that normalization is computed separately for each subtopic. Each subtopic requires its own ideal ranked list, but these lists can be computed directly. These

Table 4 Discount vectors used in evaluation measures

1. **ERR**$(k) = \frac{1}{k}$
2. **DCG**$(k) = \frac{1}{\log(k+1)}$
3. **RBP**$(k) = \frac{1}{\beta}^{k-1}$

measures have a tendency to reward systems that ignore minor subtopics (Clarke et al. 2011; Sakai et al. 2010).

4. **D#-measures** (Sakai 2012; Sakai et al. 2010; Sakai and Song 2011): D#-measures incorporate subtopic weighting into the gain function, which the authors refer to as *global gain*.

$$\text{GlobalGain}(k) = \sum_{i=1}^{M} w_i g_i^k. \tag{9}$$

This definition can be derived from the probabilistic ranking principle by assuming that

(a) intents are mutually exclusive, *i.e.* $\sum_{i=1}^{M} w_i = 1$, and
(b) the binary probability of relevance is proportional to the relevance grade.

Global gains are computed for each subtopic and normalized with regard to a single ideal ranked list. Unlike Cascade measures, this ideal ranked list can be found by a simple greedy algorithm. Measures using global gain are called D-measures. To increase the correlation with subtopic coverage, D#-measures are D-measures combined linearly with S-Recall. The mixture is controlled by a parameter $\lambda \in [0, 1]$, with $\lambda = 1$ being equivalent to pure S-Recall.

6.2 α#-IA measures

Consider a hypothetical user whose search needs are satisfied by any document relevant to their intent. If the probability of each subtopic is uniform, then S-Recall represents the percentage of such users that would be satisfied by a particular ranked list. Intent aware measures can be thought of as extending this idea to non-trivial user models and non-uniform subtopic probabilities; in this framework, intent aware measures represent the expected satisfaction of a user over all possible intents. This is an attractive model of diversity, but there is no explicit novelty component: systems will be rewarded for finding multiple documents relevant to a subtopic rather than being penalized. Cascade measures do model novelty, but they do not have this feature of explicitly averaging over intents—in a sense, they "macro-average" subtopics, whereas in this work we wish to "micro-average" them. Merging cascade measures with intent-aware measures creates a new family of intent aware cascade measures e.g. α-nDCG-IA. This family computes gains in the style of cascade measures using Eq. 7, but separately for each subtopic, with each normalized against a ranked list ideal *for that subtopic*. These separate evaluations can then be merged in the style of the intent aware measures. Unfortunately, this re-inherits the problem of rewarding systems for ignoring minor intents. Therefore, our final family has a #-measure component as well. Intent-Aware cascade #-measures are defined as a linear combination of S-Recall and an intent aware cascade measure. For example,

$$\alpha\#\text{-nDCG-IA@}k = \lambda \times \text{S-Recall@}k + (1-\lambda)\sum_{i=1}^{M} w_i \times \alpha\text{-nDCG}_i\text{@}k \tag{10}$$

6.3 Averaging over topics and subtopics

Our first goal has been to develop evaluation measures that explicitly take into account the diversity present in the collection. Our hypothesis is that all systems will perform well on the easier topics for which any ranked list is likely to be diverse, and the easier, more represented subtopics that any ranked list is likely to cover. If a system performs well with regard to these topics and subtopics, it does not provide us with much information. Therefore, we wish to place more emphasis on the more difficult topics and less prevalent subtopics, as only high quality systems should be able to perform well on these.

We focus on the difficult topics and subtopics in two ways. The first, inspired by GMAP (Robertson 2006), is by using the geometric mean. This has the effect of amplifying the impact of topics and subtopics for which a system performed poorly. By assumption, these must have been the more difficult topics and subtopics. The second is to explicitly account for the difficulty using diversity difficulty and subtopics miss rate. Since *dd* is a number between zero and one, with zero representing a topic with no diversity, we weight each topic by one minus its diversity difficulty. *smr* can be used directly.

Experimentally, we investigate the following three methodologies for averaging evaluations over topics.

1. **Avg**: the arithmetic average over the topics.
2. **Geom**: the geometric mean over the topics.
3. **DD**: *dd*-weighted average.

We also investigate the following four methodologies for averaging over subtopics.

1. **Cascade**: we do not average over subtopics. As in cascade and D# measures, ranked lists are normalized by a single ideal ranked list.
2. **Micro**: all subtopics are considered equally. This is the arithmetic average of each subtopic normalized independently.
3. **Geom**: the geometric, rather then the arithmetic, mean.
4. **SMR**: each subtopic is weighted by its subtopic miss rate.

6.4 Impact on evaluation

In this section, we explore the extent to which α#-IA measures evaluate systems differently than existing measures. Given a collection, an evaluation measure induces an ordering on the submitted runs. We use Kendall's τ (Kendall 1938) to assess the degree of correlation between the ranking of systems by different measures. By evaluating all systems submitted to TREC 2010 and 2011, we can compare the relative system rankings as computed by ERR-IA and D#-nDCG, the primary measures by which systems are evaluated at TREC and NTCIR respectively, and α#-IA measures to see how correlated they are. We can also measure the impact of the topic averaging methodologies of Sect. 6.3 by using them with ERR-IA and D#-nDCG. In our evaluations, we use the default parameters of TREC and NTCIR, and set $\alpha = 0.3$ and $\lambda = 0.5$ in α#-IA measures.[5] This can tell us whether two measures evaluate systems similarly. However, if two measures are

[5] In Sect. 6.5, these values will be shown to produce metrics with high discriminatory power.

Table 5 TREC and NTCIR gold standard vs a small sample of α#-IA measures

	ERR-IA	D#-nDCG	DD-Geom	Geom-Micro	Avg-SMR
ERR-IA	–	0.82 / 0.71	0.15 / 0.23	0.72 / 0.68	0.80 / 0.73
D#-nDCG		–	0.19 / 0.13	0.74 / 0.86	0.89 / 0.86
DD-Geom			–	0.21 / 0.16	0.20 / 0.17
Geom-Micro				–	0.79 / 0.86
Avg-SMR					–

Kendall's τ 2010 / 2011

found to be different, it cannot tell us which of the two is better. This question will be addressed in Sects. 6.5 and 7.2.

Table 5 shows the "gold standard" TREC and NTCIR measures compared with each other and several α#-IA measures. Each cell in the table shows the Kendall's τ value in 2010 and 2011, separated by a slash. With τ values ranging from roughly 0.7 to 0.9, we can see that the ERR-IA, D#-nDCG, and α#-IA measures with geometrice topic averaging and arithmetic subtopic averaging (Geom-Micro), and arithmetic topic averaging and subtopic miss rate-weighted subtopic averaging (AVG-SMR) all rank systems in highly correlated orders. However, when comparing any of these measures to α#-IA with diversity difficulty topic averaging and geometric subtopic averaging (DD-Geom), we get highly uncorrelated rankings with τ values approximately between 0.15 and 0.2. This tells us that DD-Geom evaluates systems very differently from other measures.

Tables 6 and 7 show the impact of topic averaging on the gold standard measures. With τ values ranging from 0.74 to 0.89, we can see that ordering systems by arithmetic (avg) and geometric (geom) averaging produce similar lists. However, averaging topics by their diversity difficulty (DD) produces orderings that rank systems very differently. In fact, in 2010, the τ values of 0.06 and 0.08 show that the results using ERR-IA with DD averaging are almost completely uncorrelated with the results using ERR-IA with arithmetic and geometric averaging.

Table 8 shows the impact of subtopic averaging on α#-IA measures. We observe that, independent of topic average, cascade normalization (casc) and geometric subtopic averaging (geom) are quite similar. This matches our intuition of α-nDCG (recall that arithmetic average and cascading subtopic average is a linear combination of S-Recall and α-ndcg), namely that penalizing redundancy increases the impact of difficult subtopics.

We can also see that, with a minimum τ value of 0.72, if we use arithmetic topic averaging (avg, top table), the choice of subtopic averaging does not have a large impact. The impact is somewhat larger with geometric topic averaging (geom, middle table), with a minimum of 0.6. When we use diversity difficulty (DD, bottom table) topic weighting, however, the difference becomes more dramatic, with a minimum of 0.4. However, we observe that subtopic miss rate weighting (smr) is highly similar to the arithmetic average of subtopics (micro).

That diversity difficulty averaging is so different from the other averages supports our hypothesis that evaluation should use information about the collection to emphasize difficult topics. We have shown that doing so causes evaluation metrics to prefer different systems (although we have not yet shown that it causes metrics to prefer more diverse systems). However, Table 8 also shows that subtopic miss rate-weighted averaging (smr) is very similar to the arithmetic average of subtopics (micro), suggesting that it would be better to emphasize subtopics on which systems performed poorly, rather than subtopics

Table 6 Impact of topic averaging on ERR-IA

ERR-IA	avg	geom	DD
avg	–	0.89 / 0.81	0.06 / 0.25
geom		–	0.08 / 0.17
DD			–

Kendall's τ 2010 / 2011

Table 7 Impact of topic averaging on D#-nDCG

D#-nDCG	avg	geom	DD
avg	–	0.74 / 0.88	0.50 / 0.29
geom		–	0.56 / 0.34
DD			–

Kendall's τ 2010 / 2011

Table 8 Impact of subtopic averaging on α#-IA measures

	casc	micro	geom	smr
avg				
casc	–	0.87 / 0.77	0.94 / 0.87	0.81 / 0.73
micro		–	0.82 / 0.74	0.90 / 0.94
geom			–	0.80 / 0.72
smr				–
geom				
casc	–	0.72 / 0.66	1.00 / 0.92	0.72 / 0.65
micro		–	0.73 / 0.62	0.99 / 0.97
geom			–	0.72 / 0.60
smr				–
DD				
casc	–	0.44 / 0.63	0.99 / 1.00	0.40 / 0.60
micro		–	0.45 / 0.63	0.90 / 0.94
geom			–	0.41 / 0.60
smr				–

Kendall's τ 2010 / 2011

that we expect to be difficult. We believe that our hypothesis is valid, but our approximation of *smr* is not. We discuss this further in Sect. 7.2 and in our conclusion.

6.5 Disciminative power

We have shown that our measures can rank systems differently than other diversity measures. In this section, we show that this does not sacrifice the sensitivity of evaluation to changes in ranked lists. One of the primary measures of sensitivity appearing in the IR literature (Clarke et al. 2011; Sakai 2012; Sakai et al. 2010; Sakai and Song 2011) is discriminative power (Sakai 2006). Discriminative power measures sensitivity by conducting statistical significance testing on pairs of systems. Given the same set of queries, two different systems will produce different ranked lists. Ideally, measures should produce different sets of evaluations. The discriminative power of a measure is defined as the percentage of system pairs that are significantly different.

In this section, we compare the discriminative power of α#-IA measures with that of existing measures. There are, essentially, three aspects of α#-IA measures that can be varied: our choice of discount function, our tuning of the α and λ parameters, and our choice of topic and subtopic normalization. Unfortunately, it is not immediately obvious how to measure the effect of topic averaging on discriminative power. We leave this for future work. In this section, we focus on subtopic normalization. In all experiments, α—which models a user's tolerance for redundancy—and λ—which controls the mixture with S-Recall—vary over the set $\{0, 0.1, 0.2, \ldots, 1\}$. Following Clarke et al. (2011), when using RBP, β is set to 0.8. Discriminative power experiments use a two-tailed paired *t* test. The tests are bootstrapped (Sakai 2006), with $B = 1{,}000$.

Table 9 shows the maximum discriminative power at rank 20 of each α#-IA measure observed as α and λ are varied, as well as the maximum observed value of the D# measures as λ is varied. From this table we observe that no measures have substantially more discriminatory power than any other when parameters are appropriately tuned. We note that the α#-IA measures have more discriminatory power than D# measures, though not substantially.

Figure 3 shows the discriminative power of each evaluation measure at rank 20 with DCG discounting for all values of α and λ. These results were found to be typical for all three discount functions. We can compare the α#-IA measures to existing measures (with the exception of D# measures) by carefully considering these plots. For any subtopic average, setting $\lambda = 1$ (the far-right side in 3D plots) shows S-Recall. Using the cascade average and setting $\lambda = 0$ (the near-left side in 3D plots) shows α-nDCG. If you assume that the subtopic distribution is uniform, then using the arithmetic average (micro) and setting $\lambda = \alpha = 0$ (the leftmost corner) shows nDCG-IA. Since the maximum for each year is achieved by cascade averaging, and not on the near-left or far-right side (i.e. it is achieved with $0 < \lambda < 1$), we can conclude that the α#-IA measures do have somewhat higher discriminatory power than existing measures.

Table 9 Maximum discriminative power observed on actual runs at rank 20

Discount	Subtopic	2010 Max	2011 Max
ERR	Cascade	0.677	0.606
	Micro	0.673	0.610
	Geom	0.627	0.593
	SMR	0.659	0.601
D#-ERR		0.667	0.607
DCG	Cascade	0.675	0.617
	Micro	0.673	0.623
	Geom	0.617	0.595
	SMR	0.651	0.607
D#-nDCG		0.643	0.608
RBP	Cascade	0.677	0.607
	Micro	0.677	0.621
	Geom	0.617	0.595
	SMR	0.657	0.600
D#-RBP		0.653	0.595

All choices of subtopic average and discount function have comparable maxima

From Fig. 3, we observe that setting $\alpha = 0.3$ and $\lambda = 0.5$ seem to be reasonable choices to use in further investigation. Figures 4 and 5 show all four subtopic averages at ranks 5, 10, and 20 as one of the parameters is fixed while the other is allowed to vary. From these we conclude that while the subtopic averages that emphasize the difficult subtopics—the geometric average (geom) and the subtopic miss rate-weighted average (smr)—have lower discriminative power overall, they are comparable when α and λ are appropriately tuned.

The results of this section as a whole tell us that α#-IA measures are slightly more sensitive than existing measures, as assessed by discriminative power.

7 Document selection sensitivity

A system's diversity is necessarily conflated with the diversity of the collection and the system's ad-hoc performance. So far, we have described ways of incorporating collection diversity into evaluation metrics. In this section, we introduce **Document selection**

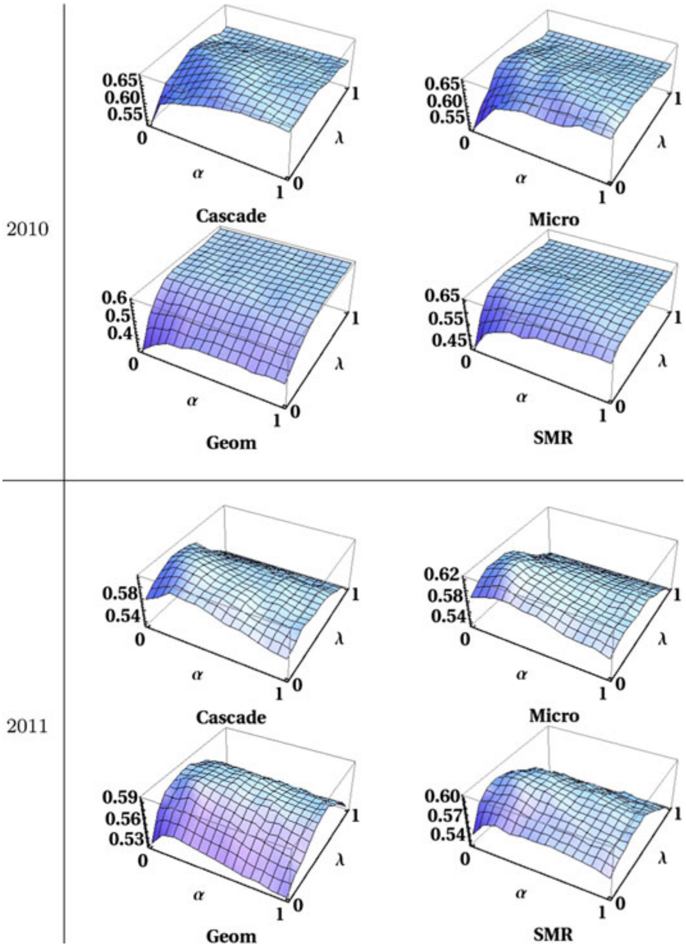


Fig. 3 Discriminative power as a function of α and λ

sensitivity to address the issue of controlling for ad-hoc performance. Document selection sensitivity is defined in Sect. 7.1. Experimental results are presented in Sect. 7.2.

7.1 Definition

One way to control for the impact of ad-hoc performance on diversity evaluation is to only consider lists with the same performance. We do so using artificial ranked lists created by randomly permuting the set of relevant documents. These lists will all have the same ad-hoc performance—perfect. For a given topic, whatever difference exists between ranked lists with perfect combined precision—the percentage of retrieved documents relevant to at least one subtopic (Clarke et al. 2011)—must be due solely to novelty and diversity.

We use the binary TREC relevance judgments to create 50 artificial runs by (uniformly) randomly ranking the relevant documents of all 100 TREC topics. Figure 6 shows the discriminative power of α#-IA measures over these simulated runs. We observe that discriminative power is almost an order of magnitude smaller than it is over actual runs. This suggests that diversity evaluation is more a function of ad-hoc performance than of actual novelty and diversity. Given this, if we wish to assess the extent to which measures are impacted by novelty and diversity, we must use an alternative framework.

For some measure M, imagine evaluating randomly selected permutations of the set of relevant documents. Since our measures are truncated at specific ranks, these lists will, effectively, differ in the relevant documents selected, as well as the ranks of the relevant documents. The observed set of evaluations has some sample mean, $\bar{x}$, and sample standard deviation, s. We define the **document selection sensitivity** of a measure as the coefficient of variation—the standard deviation divided by the mean—of the set of evaluations.

$$dss(M) = \frac{s}{\bar{x}} \tag{11}$$

This produces a normalized measure of the variance of ranked lists with perfect performance. A low document selection sensitivity means that it is unlikely that a system which

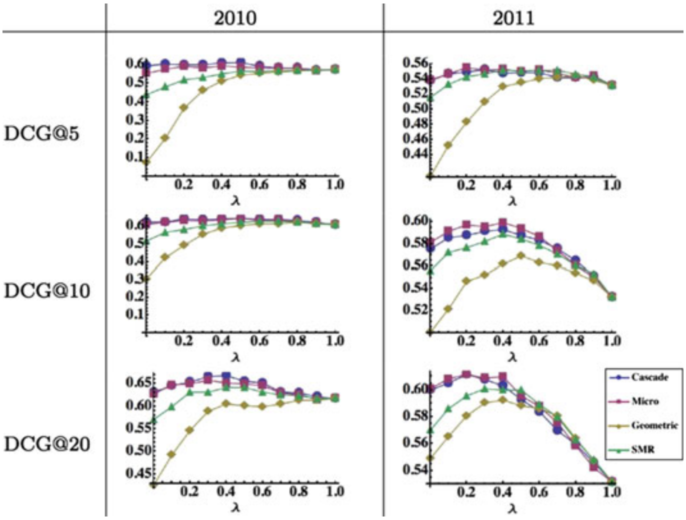


Fig. 4 Discriminative power of as a function of λ. α is fixed at 0.3. While choice of subtopic average clearly impacts discriminative power, all maxima are comparable

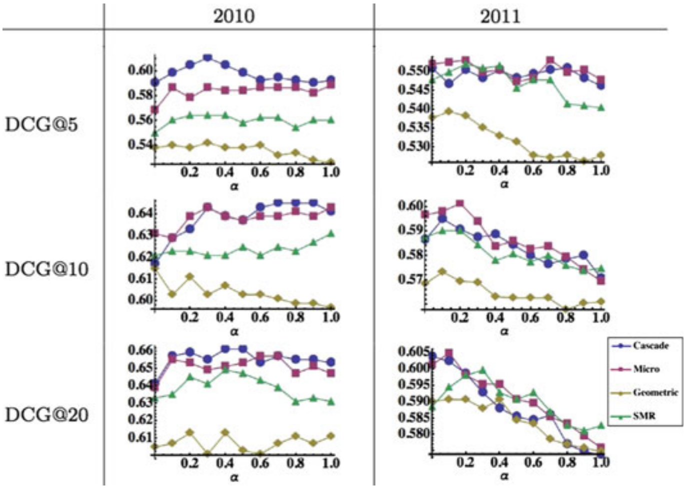
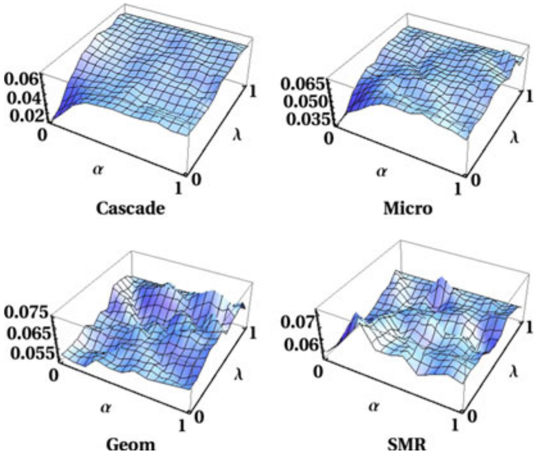


Fig. 5 Discriminative power as a function of α. λ is fixed at 0.5. While choice of subtopic average clearly impacts discriminative power, all maxima are comparable

Fig. 6 Discriminative power on artificial ranked lists with perfect combined precision. Discriminative power is almost an order of magnitude smaller on these lists than on actual systems



assigns relevant documents at random will be different than the mean.[6] The larger this number, the more sensitive the measure is to the documents in ranked lists and their order. If this number is small, it does not mean that the measure did not evaluate some lists as substantially better than others: the maximum and minimum scores achieved by any of the randomly generated ranked lists could be quite different; it simply means that the majority of ranked lists have highly similar scores.

Document selection sensitivity is computed separately for each topic. As discussed in Sect. 6.3, when averaging, we wish to place more emphasis on difficult topics with low

[6] For normally-distributed data, a coefficient of variation of c % means that roughly 68 % of the population is within ±c % of the mean.

diversity difficulty scores. Experimentally, we explore the impact of the three topic averages previously discussed: Avg, Geom, and DD.

We note that as described, document selection sensitivity does not use relevance grades or intent probabilities. We suggest that one way to use this information would be to select documents iteratively by first using the intent probability distribution to choose a subtopic, and then randomly drawing from the most relevant documents remaining for that subtopic. We hypothesize that not using relevance grades and subtopic probabilities may be beneficial, in that some diversity measures will be better able to leverage this information to distinguish between lists than others. We will address this question in future work.

A further limitation is that, as defined, DSS can only be used to evaluate measures that are always positive. While this is almost always the case, it is not universal, e.g. logit(AP). Also, DSS is not invariant to even the most simple of transformations. For example, given a measure M, the measure $M' = \frac{M+1}{2}$ would have a smaller DSS score, while still ranking systems in the exact same order.

7.2 Experimental results

For each topic, 1,000 ranked lists were created by ranking the relevant documents at uniformly random. These ranked lists were used to compute the document selection sensitivity of each measure on each topic. As before, α—which models a user's tolerance for redundancy—and λ—which controls the mixture with S-Recall—vary over the set $\{0, 0.1, 0.2, \ldots, 1\}$. Table 10 shows that, unlike discriminatory power, document selection sensitivity can be affected by choice of topic and subtopic averaging. At rank 20, document selection sensitivity ranges from a low of 0.05 to a high of 0.8. This can also be seen in Fig. 7, which shows the selection sensitivity using DCG discounting at rank $k = 20$. Geometric (geom) and arithmetic topic averaging are quite similar, and the latter is omitted. Diversity difficulty topic weighting (DD) shows marked increases in selection sensitivity, as does geometric subtopic weighting (geom). Subtopic miss rate weighting (smr) has higher selection sensitivity than arithmetic average (micro) and cascade normalization.

Figures 8 and 9 show the selection sensitivity of the topic averages at ranks $k = 5$, 10, and 20, each with one of the parameters fixed. From these figures, as well as Table 10, we can see that selection sensitivity clearly goes down as the cutoff increases. This makes sense intuitively. Imagine that there are 20 relevant documents, 5 documents relevant to all subtopics and 15 documents each relevant to a single subtopics. At $k = 20$, you will get at least some gain from every relevant document. At $k = 5$, you may see the 5 documents relevant to all subtopics, or you may see none of them. Seeing all of them versus none of them should have more variance than seeing all of them in different orders.

Consulting Table 10, we can see that α#-IA measures clearly have higher document selection sensitivity than D# measures. We can compare the α#-IA measures against the other existing measures by carefully considering Fig. 7. Again, for any subtopic average, setting $\lambda = 1$ (the far-right side in 3D plots) shows S-Recall. Using the cascade average and setting $\lambda = 0$ (the near-left side in 3D plots) shows α-nDCG. If you assume that the subtopic distribution is uniform, then using the arithmetic subtopic averaging (micro) and setting $\lambda = \alpha = 0$ (the leftmost corner) shows nDCG-IA. Since the maximum is achieved by geometric subtopic averaging (geom), and not on the far-right side where $\lambda = 1$, we can conclude that the α#-IA measures can have significantly higher document selection sensitivity than existing measures.

Table 10 Maximum observed document selection sensitivity

Topic	Subtopic	Rank	Max		
			ERR	DCG	RBP
Avg	Cascade	5	0.229	0.218	0.218
		10	0.210	0.174	0.176
		20	0.204	0.152	0.170
	Micro	5	0.229	0.218	0.218
		10	0.211	0.174	0.174
		20	0.204	0.154	0.154
	Geom	5	1.558	1.556	1.556
		10	1.012	0.978	0.978
		20	0.683	0.608	0.608
	SMR	5	0.598	0.529	0.516
		10	0.518	0.404	0.422
		20	0.480	0.327	0.391
	D#	5	0.218	0.218	0.218
		10	0.152	0.152	0.152
		20	0.116	0.096	0.111
Geom	Cascade	5	0.224	0.204	0.196
		10	0.205	0.168	0.167
		20	0.199	0.149	0.162
	Micro	5	0.224	0.204	0.204
		10	0.206	0.169	0.169
		20	0.199	0.151	0.151
	Geom	5	1.176	1.120	1.120
		10	0.761	0.667	0.667
		20	0.540	0.420	0.420
	SMR	5	0.504	0.436	0.423
		10	0.448	0.344	0.344
		20	0.425	0.292	0.323
	D#	5	0.176	0.175	0.175
		10	0.109	0.095	0.097
		20	0.081	0.057	0.077
DD	Cascade	5	0.260	0.260	0.260
		10	0.220	0.194	0.201
		20	0.211	0.163	0.193
	Micro	5	0.260	0.260	0.260
		10	0.220	0.194	0.194
		20	0.211	0.165	0.165
	Geom	5	2.015	2.029	2.029
		10	1.319	1.300	1.300
		20	0.871	0.801	0.801
	SMR	5	0.719	0.642	0.629
		10	0.612	0.492	0.518
		20	0.559	0.383	0.478
	D#	5	0.260	0.260	0.260
		10	0.194	0.194	0.194
		20	0.129	0.129	0.129

Smaller cutoffs have higher sensitivity

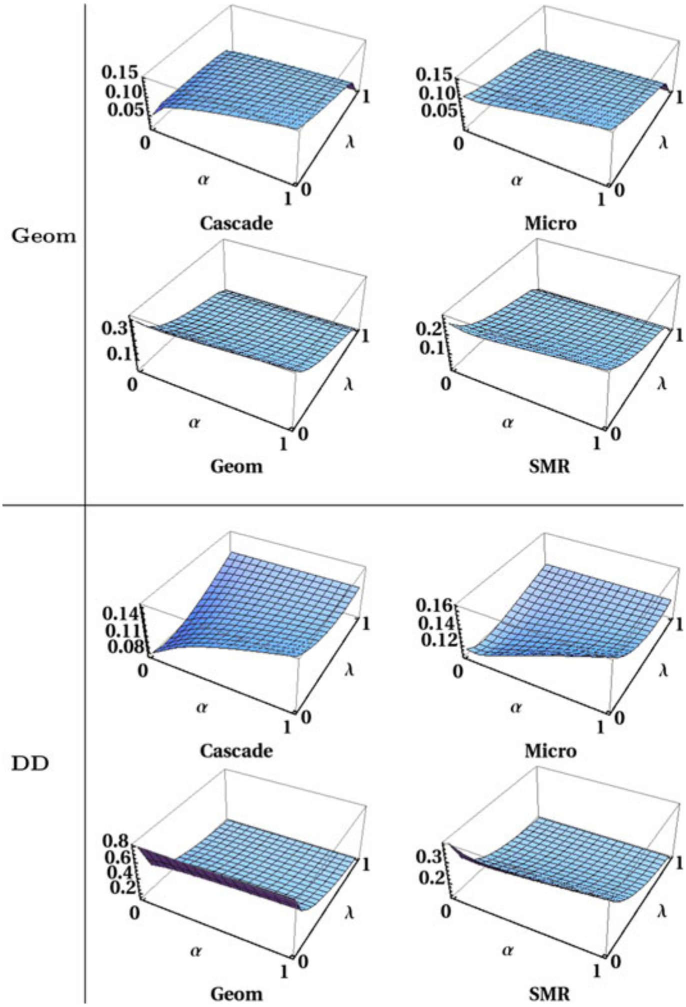


Fig. 7 Document selection sensitivity as a function of α and λ. Choice of topic and subtopic averaging can have a substantial impact

According to document selection sensitivity, one should use diversity difficulty topic averaging and geometric subtopic averaging. That one should use diversity difficulty topic averaging supports our hypothesis that evaluation should take a collection-oriented view, emphasizing topics that are difficult, rather than a user-oriented view, emphasizing topics with poor results. However, as with Sect. 6.4, geometric subtopic averaging outperforms subtopic miss rate-weighted averaging. This would seem to directly contradict our hypothesis; by emphasizing small values, geometric subtopic averaging takes the user-oriented view, emphasizing subtopics on which the user is expected to be left unsatisfied. Instead, we believe that this is likely due to our particular implementation of subtopic miss rate, both the approximation of sampling with replacement, and by choosing to measure subtopic miss rate at ξ, which is often as high as rank one or two. We believe that geometric subtopic averaging, by emphasizing the subtopics where systems performed poorly, as would be expected of

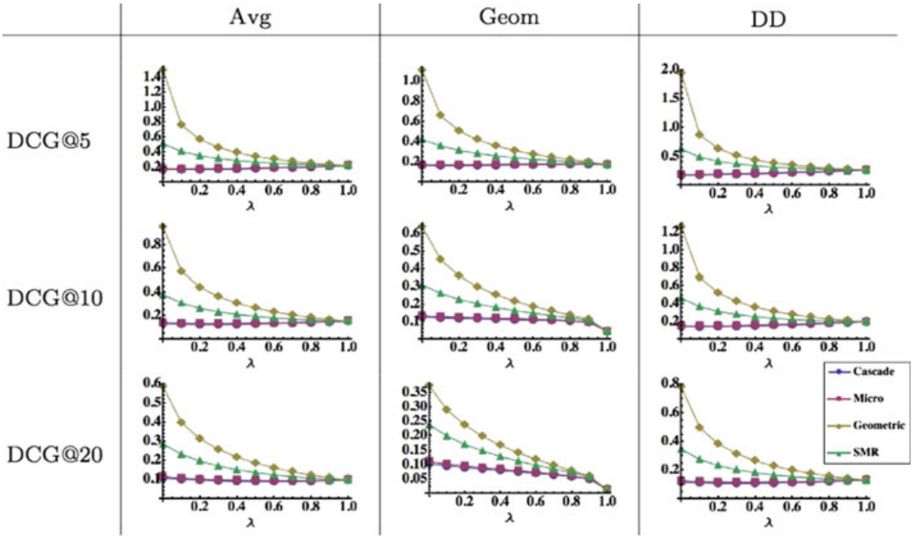


Fig. 8 Document selection sensitivity as a function of λ. α is fixed at 0.3

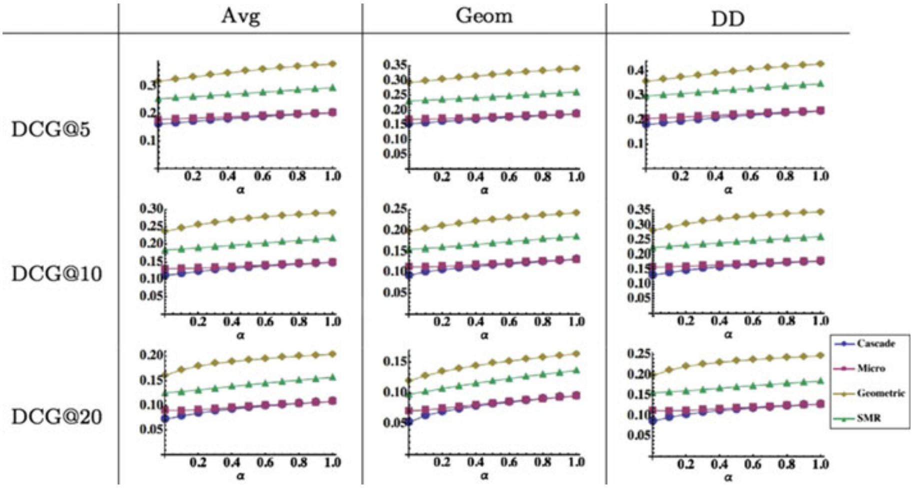


Fig. 9 Document selection sensitivity as a function of α. λ is fixed at 0.5

subtopics with large subtopic miss rates, was actually a better approximation of subtopic miss rate than our computed approximation. We will revisit this in future work. We do observe that *smr*, even as approximated here, still outperforms micro and cascade normalization with respect to document selection sensitivity.

8 Conclusion and future work

A search engine's diversity is necessarily conflated with its ability to perform ad-hoc retrieval and the diversity of the collection. In this work, we attempted to isolate diversity from those other factors so that we can begin to understand it. We (1) introduced a new family of measures that explicitly account for the collection diversity and (2) introduced a meta-evaluation measure of sensitivity that controls for ad-hoc performance. Our hypothesis is that these collection-oriented features, while opaque to the user, will be better able to differentiate between systems, thereby leading to a better overall user experience.

To assess collection difficulty, we developed measures at the topic and subtopic level. At the topic level, diversity difficulty blends the maximum possible number of subtopics covered by any ranked list with the number of subtopics covered by the expected ranked list. At the subtopic level, subtopic miss rate measures the probability of selecting documents at random and failing to cover subtopics. We showed that α#-IA measures, which combine the best features of existing evaluation measures and emphasize difficult topics and subtopics, sometimes rank systems in quite different orders than existing measures, yet have slightly more discriminative power.

That our measures prefer different systems does not indicate that they prefer more diverse systems. To show that our new measures preferred more diverse systems than existing measures, we restricted our attention to artificial ranked lists with perfect combined precision to show that our measures were less influenced by ad-hoc performance than existing measures. According to discriminative power, no measure was able to distinguish between these lists. This led us to introduce document selection sensitivity, the coefficient of variation of an evaluation measure over these artificial ranked lists. According to this measure, α#-IA measures that explicitly account for collection diversity were far more sensitive to differences in these lists than existing measures, suggesting that these measures may prefer more diverse systems. However, while averaging subtopics by their difficulty also led to higher document selection sensitivity, it was still less than geometric averaging. This is likely due to limitations of our implementation of difficulty at the subtopic level.

We believe that these results support our hypothesis that taking a collection-oriented view of evaluation leads to systems that are preferable to the user. We contrast this with the user-oriented view of Sakai's intuitiveness measure (Sakai 2012; Sakai et al. 2010; Sakai and Song 2011). We look forward to comparing these two approaches, in terms of correlation with each other and with the preference of actual users.

We note that our framework accepts any definition of difficulty at the collection level. In future work, we will explore alternate definitions of, and uses for, diversity difficulty at the topic and subtopic levels. We also wish to explore the correlation with diversity difficulty and ad-hoc query difficulty. Is one predictive of the other?

There is also the question of incorporating relevance grades and intent probabilities into document selection sensitivity. We briefly suggested one way this can be done, but surely there are other ways. Would incorporating this information produce a more useful meta-evaluation?

Finally, recent work has shown that subtopic taxonomy (Broder 2002), e.g. whether the subtopic is *navigational* or *informational*, has been shown to lead to better performance of both diversification algorithms (Santos et al. 2011) and diversity evaluation measures (Sakai 2012), since a user is far less tolerant of redundancy for a navigational query than an informational one. In future work, we intend to show the effect of incorporating subtopic taxonomy into document selection sensitivity and α#-IA measures.

References

Agrawal, R., Gollapudi, S., Halverson, A., Ieong, S. (2009). Diversifying search results. In *Proceedings of the second ACM international conference on web search and data mining, WSDM '09* (pp. 5–14). New York, NY: ACM.

Aslam, J. A., & Pavlu, V. (2007). Query hardness estimation using jensen-shannon divergence among multiple scoring functions. In *Proceedings of the 29th European conference on IR research, ECIR'07* (pp. 198–209). Berlin: Springer.

Broder, A. (2002). A taxonomy of web search. *SIGIR Forum, 36*(2), 3–10.

Carbonell, J., & Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '98* (pp. 335–336). New York, NY: ACM.

Carmel, D., & Yom-Tov, E. (2010). *Estimating the query difficulty for information retrieval. Synthesis lectures on information concepts, retrieval, and services*. San Rafael: Morgan & Claypool.

Carmel, D., Yom-Tov, E., Darlow, A., Pelleg, D. (2006). What makes a query difficult? In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '06* (pp. 390–397). New York, NY: ACM.

Carterette, B. (2009). An analysis of NP-completeness in novelty and diversity ranking. In *Proceedings of the 2nd international conference on theory of information retrieval: Advances in information retrieval theory, ICTIR '09* (pp. 200–211). Berlin : Springer.

Chapelle, O., Metlzer, D., Zhang, Y., Grinspan, P. (2009). Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM conference on information and knowledge management, CIKM '09* (pp. 621–630). New York, NY: ACM.

Chen, H., & Karger, D. R. (2006) Less is more: Probabilistic models for retrieving fewer relevant documents. In *Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '06* (pp. 429–436). New York, NY: ACM.

Clarke, C. L., Craswell, N., Soboroff, I., Ashkan, A. (2011). A comparative analysis of cascade measures for novelty and diversity. In *Proceedings of the fourth ACM international conference on web search and data mining, WSDM '11* (pp. 75–84). New York, NY: ACM.

Clarke, C. L., Kolla, M., Cormack, G. V., Vechtomova, O., Ashkan, A., Büttcher, S., MacKinnon, I. (2008). Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '08* (pp. 659–666). New York, NY: ACM.

Clarke, C. L., Kolla, M., Vechtomova, O. (2009). An effectiveness measure for ambiguous and underspecified queries. In *Proceedings of the 2nd international conference on theory of information retrieval: Advances in information retrieval theory, ICTIR '09* (pp. 188–199). Berlin: Springer.

Clarke, C. L. A., Craswell, N., Soboroff, I. (2009). Overview of the TREC 2009 web track. In *18th text retrieval conference*. Maryland: Gaithersburg.

Clarke, C. L. A., Craswell, N., Soboroff, I., Cormack, G. V. (2010). Overview of the TREC 2010 web track. In *19th text retrieval conference*. Maryland: Gaithersburg.

Clarke, C. L. A., Craswell, N., Soboroff, I., Voorhees, E. M. (2011). Overview of the TREC 2011 web track. In *20th text retrieval conference*. Maryland: Gaithersburg.

Craswell, N., Zoeter, O., Taylor, M., Ramsey, B. (2008). An experimental comparison of click position-bias models. In *Proceedings of the international conference on web search and data mining, WSDM '08* (pp. 87–94). New York, NY: ACM.

Cronen-Townsend, S., Zhou, Y., Croft, W. B. (2002). Predicting query performance. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '02* (pp. 299–306). New York, NY: ACM.

Cronen-Townsend, S., Zhou, Y., Croft, W. B. (2006). Precision prediction based on ranked list coherence. *Information Retrieval 9*(6), 723–755. http://link.springer.com/article/10.1007%2Fs10791-006-9006-4.

Golbus, P. B., Pavlu, V., Aslam, J. A. (2012) What we talk about when we talk about diversity. In *Proceedings of diversity in document retrieval 2012*.

Hauff, C. (2010). *Predicting the effectiveness of queries and retrieval systems*. Ph.D. thesis, University of Twente, Enschede.

He, B., & Ounis, I. (2004). Inferring query performance using pre-retrieval predictors. In *Proceedings of symposium on string processing and information retrieval* (pp. 43–54). Berlin: Springer.

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems 20*(4), 422–446.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika 30*(1/2), 81–93.

Moffat, A., & Zobel, J. (2008). Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems 27*(1), 2:1–2:27.

Mothe, J., & Tanguy, L. (2005). Linguistic features to predict query difficulty. In *In ACM SIGIR 2005 workshop on predicting query difficulty—methods and applications.*

Robertson, S. (2006). On GMAP: And other transformations. In *Proceedings of the 15th ACM international conference on information and knowledge management, CIKM '06* (pp. 78–83). New York, NY: ACM.

Robertson, S. E. (1977). The probability ranking principle in IR. *Journal of Documentation 33*(4), 294–304.

Sakai, T. (2006). Evaluating evaluation metrics based on the bootstrap. In *Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '06* (pp. 525–532). New York, NY: ACM.

Sakai T. (2012). Evaluation with informational and navigational intents. In *Proceedings of the 21st world wide web conference (WWW) 2012*.

Sakai, T., Craswell, N., Song, R., Robertson, S., Dou, Z., Lin, C. Y. (2010). Simple evaluation metrics for diversified search results. In *The Third international workshop on evaluating information access (EVIA)*.

Sakai, T., & Song, R. (2011). Evaluating diversified search results using per-intent graded relevance. In *Proceedings of the 34th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '11* (pp. 1043–1052). New York, NY: ACM.

Santos, R., Macdonald, C., Ounis, I. (2012). On the role of novelty for search result diversification. *Information Retrieval 15*, 478–502.

Santos, R. L., Macdonald, C., Ounis, I. (2011). Intent-aware search result diversification. In *Proceedings of the 34th international ACM SIGIR conference on research and development in information retrieval, SIGIR '11* (pp. 595–604). New York, NY: ACM.

Shtok, A., Kurland, O., Carmel, D. (2009). Predicting query performance by query-drift estimation. In *Proceedings of the 2nd international conference on theory of information retrieval: Advances in information retrieval theory, ICTIR '09* (pp. 305–312). Berlin: Springer.

Vargas, S., Castells, P., Vallet, D. (2012). Explicit relevance models in intent-oriented information retrieval diversification. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, SIGIR '12* (pp. 75–84). New York, NY: ACM.

Vinay, V., Cox, I. J., Milic-Frayling, N., Wood, K. (2006). On ranking the effectiveness of searches. In *Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '06* (pp. 398–404). New York, NY: ACM.

Yilmaz, E., Shokouhi, M., Craswell, N., Robertson, S. (2010). Expected browsing utility for web search evaluation. In *Proceedings of the 19th ACM international conference on information and knowledge management, CIKM '10* (pp. 1561–1564). New York, NY: ACM.

Yom-Tov, E., Fine, S., Carmel, D., Darlow, A. (2005). Metasearch and federation using query difficulty prediction. In *In ACM SIGIR 2005 workshop on predicting query difficulty—methods and applications*.

Zhai, C. X., Cohen, W. W., Lafferty, J. (2003). Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. In *Proceedings of the 26th annual international ACM SIGIR conference on research and development in informaion retrieval, SIGIR '03* (pp. 10–17). New York, NY: ACM.

Zhang, Y., Park, L., Moffat, A. (2010). Click-based evidence for decaying weight distributions in search effectiveness metrics. *Information Retrieval 13*, 46–69. doi:10.1007/s10791-009-9099-7.

Zhou, Y., & Croft, W. B. (2006). Ranking robustness: A novel framework to predict query performance. In *Proceedings of the 15th ACM international conference on information and knowledge management, CIKM '06* (pp. 567–574). New York, NY: ACM.

Zhou, Y., & Croft, W. B. (2007). Query performance prediction in web search environments. In *Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '07* (pp. 543–550). New York, NY: ACM.

Zuccon, G., Azzopardi, L., Zhang, D., Wang, J. (2012). Top-k retrieval using facility location analysis. In *Proceedings of the 34th European conference on advances in information retrieval, ECIR'12* (pp. 305–316). Berlin: Springer.