A Comparative Analysis of Cascade Measures for Novelty and Diversity

Charles L. A. Clarke
University of Waterloo

Nick Craswell
Microsoft

Ian Soboroff
NIST

Azin Ashkan
University of Waterloo

ABSTRACT

Traditional editorial effectiveness measures, such as nDCG, remain standard for Web search evaluation. Unfortunately, these traditional measures can inappropriately reward redundant information and can fail to reflect the broad range of user needs that can underlie a Web query. To address these deficiencies, several researchers have recently proposed effectiveness measures for novelty and diversity. Many of these measures are based on simple *cascade* models of user behavior, which operate by considering the relationship between successive elements of a result list. The properties of these measures are still poorly understood, and it is not clear from prior research that they work as intended. In this paper we examine the properties and performance of cascade measures with the goal of validating them as tools for measuring effectiveness. We explore their commonalities and differences, placing them in a unified framework; we discuss their theoretical difficulties and limitations, and compare the measures experimentally, contrasting them against traditional measures and against other approaches to measuring novelty. Data collected by the TREC 2009 Web Track is used as the basis for our experimental comparison. Our results indicate that these measures reward systems that achieve an balance between novelty and overall precision in their result lists, as intended. Nonetheless, other measures provide insights not captured by the cascade measures, and we suggest that future evaluation efforts continue to report a variety of measures.

Categories and Subject Descriptors

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

General Terms

Experimentation, Measurement

Keywords

effectiveness measures, novelty, diversity

WSDM'11, February 9–12, 2011, Hong Kong, China.

1. NOVELTY AND DIVERSITY

Queries mean different things to different users. A user submitting the query "defender" to a Web search engine could be seeking information regarding any of several possible things, including the Windows Defender anti-spyware program, the Land Rover Defender sport-utility vehicle, the Williams Defender arcade game, the Chicago Defender newspaper, or the Defender marine supply company. Even when users share a common interpretation of a query, their information needs may still differ. One user may be seeking the homepage of Windows Defender in order to download the software; a second user might be interested in product reports; a third user may be interested in both.

Ideally, search engines should return results that reflect the diversity of their users' information needs. For an ambiguous query, an ideal result page would include an appropriate mixture of results that address its possible interpretations. Even when the topic of a query is unambiguous the results should include appropriate coverage of different aspects of that topic, since queries are rarely specific enough to pinpoint an information need. As the user scans the result list, she should encounter novel information with each new result. The relative popularity of subtopics should inform the order in which the search engine presents these results.

Evaluation measures for novelty and diversity attempt to quantify the extent to which a result list appropriately addresses the breadth of possible information needs underlying a query. Most proposed measures explicitly decompose these possible information needs into a number of subtopics and compute the degree to which a result list provides coverage of these subtopics. Foundational work by Zhai et al. [21] defines a number of subtopic recall measures for this purpose. Zhang et al. [22] and Zhu et al. [24] also measure diversity in terms of subtopic recall. More recently, work by Clarke et al. [7,9], Agrawal et al. [1], and Chapelle et al. [5] propose weighted linear combinations of measures computed with respect to the individual subtopics, and these proposals form the focus of our paper. Sakai et al. [17] raise concerns about these proposals, which we attempt to address. Other work measures novelty and diversity through user studies [3], implicit user feedback [13], and comparisons with Wikipedia disambiguation pages and ODP categories [14]

Traditionally, effectiveness measures for ranked retrieval are grounded in the *probability ranking principle*, which states that the overall effectiveness of a system to its users will be maximized by ranking the documents in the collection in order of decreasing probability of relevance. Under these traditional measures, such as average precision and nDCG [11],

```
<topic number="20" type="ambiguous">
  <query>defender</query>
  <description> I’m looking for the homepage of Windows Defender, an anti-spyware program.</description>
  <subtopic number="1" type="nav">
    I’m looking for the homepage of Windows Defender, an anti-spyware program. </subtopic>
  <subtopic number="2" type="inf">
    Find information on the Land Rover Defender sport-utility vehicle. </subtopic>
  <subtopic number="3" type="nav">
    I want to go to the homepage for Defender Marine Supplies. </subtopic>
  <subtopic number="4" type="inf">
    I’m looking for information on Defender, an arcade game by Williams. Is it possible to play it online? </subtopic>
  <subtopic number="5" type="inf">
    I’d like to find user reports about Windows Defender, particularly problems with the software. </subtopic>
  <subtopic number="6" type="nav">
    Take me to the homepage for the Chicago Defender newspaper. </subtopic>
</topic>

<topic number="47" type="faceted">
  <query>indexed annuity</query>
  <description>I’m looking for information about indexed annuities.  </description>
  <subtopic number="1" type="inf">
    What is an indexed annuity? What are their advantages and disadvantages? What kinds... are there? </subtopic>
  <subtopic number="2" type="inf">
    Where can I buy an indexed annuity?  What investment companies offer them? </subtopic>
  <subtopic number="3" type="inf">
    Find ratings of indexed annuities. </subtopic>
</topic>
```

Figure 1: Topics 20 and 47 from the TREC 2009 Web Track

the relevance of a retrieved document is treated independently of others in the result list. Even when the result list contains documents that are nearly identical, each is given full credit in the computation of the measure.

Considerations of novelty suggest a modification of this principle, in which we explicitly penalize redundancy in a result list by judging documents in the context of the those already seen by the user. These considerations lead to the development of effectiveness measures based on what Craswell et al. [10] call the *cascade model* of user behavior. Under these models, users are assumed to scan result lists from the top down, eventually stopping because either their information need is satisfied or their patience is exhausted [5,7,9,19, 20]. Notably, Chapelle et al. [5] demonstrate experimentally that their particular cascade measure correlates better with user behavioral metrics than do traditional measures.

To provide a forum for the experimental investigation of novelty and diversity in ranked retrieval, the TREC 2009 Web track included a new diversity task, along with its traditional ad hoc retrieval task [8]. The track organizers constructed 50 new topics for this track that include explicit subtopics for the purpose of measuring novelty and diversity.

Figure 1 presents topic number 20 and number 47 from the track. Consider topic number 20, which we used as the basis for our opening example. The query field of the topic indicates the query as entered by users. The description field was used by the adhoc task as the basis for traditional ad hoc relevance judgments. The remainder of topic number 20 comprises six subtopics, most of which represent distinct interpretations of the the query “defender”. Subtopics 1 and 5 concern different aspects of a common interpretation, related to the Windows defender program, although it is unlikely that a single page could satisfy both information needs. These subtopics were created with the aid of information extracted from the logs of a commercial search engine, and so we may view them as indicative of genuine user requirements.

Topic 20 as a whole is labeled with the attribute “ambiguous”, indicating that (for the most part) the subtopics represent distinct interpretations of the query. Other topics from the track, such as topic number 47, are labeled with the attribute “faceted”, indicating that their subtopics represent different aspects of a single common interpretation. Each subtopic is labeled with an attribute indicating whether it is navigational (“nav”) or informational (“inf”). Navigational subtopics seek a specific Web page or site, while informational subtopics seek specific content.

To generate retrieval runs, participants in the track were given only the queries associated with each topic. The full topics were not released until after all experimental results had been submitted by the participants. Using their own search systems, participants executed these queries against a common collection of Web data, known as the ClueWeb09 collection[1]. This collection was crawled from the general Web in early 2009 and contains roughly a billion documents.

After executing the queries, the participants submitted ranked lists of documents to the TREC organizers, with the 26 participating groups submitting a total of 49 experimental runs. The submissions were pooled, so that the top 20 documents from each run were judged. Hired assessors made binary relevance judgments with respect to each subtopic. These judgments were used to compute the official results, which reported various measures proposed in the papers listed above.

The proposed measures for novelty and diversity are still poorly understood, and it is not clear from prior research that they even work as intended, appropriately rewarding systems that return novel and relevant results. In the remainder of this paper we address this problem by explor-

[1] boston.lti.cs.cmu.edu/Data/clueweb09

ing theoretical and empirical properties of these measures. We use data collected by the diversity task of the TREC 2009 Web Track as the basis for our empirical evaluation. We examine common and competing ideas embedded in the proposed measures, extending and unifying them into a common framework. We highlight a major theoretical concern with some proposals — a problem that is not shared by the cascade measures. We demonstrate the need for normalization, examine competing approach to normalization, and address related concerns raised in prior work. Using simple measures of subtopic recall and traditional precision, we tease apart elements of the cascade measures, providing evidence that they combine aspects of the simpler measures. We examine correlations between the measures of novelty and diversity, and their relationship with traditional measures. Finally, we consider the discriminative power of the measures, particularly in comparison with traditional measures.

2. CASCADE MEASURES

Clarke et al. [7] suggest a useful distinction between novelty and diversity. They view diversity as a property of the information needs underlying a query, while they view novelty as a property of a retrieval result intended to address that diversity. For example, topic 20 describes the diversity underlying the query "defender" without reference to a specific result list. A certain proportion of the users submitting this query to a search engine will be interested in each of the subtopics. Novelty reflects the degree to which a specific result list accommodates this user population. Different search engines might attempt to address the variety of information needs underlying the query by returning different sets of Web pages in different orders.

Both Clarke et al. [7] and Chapelle et al. [5] suggest measuring novelty independently for each subtopic and then combining individual subtopic scores into a single overall score according to the diversity underlying the query. To achieve the aim of measuring novelty independently for each subtopic, they penalize redundancy through the application of a cascade model, rather than directly rewarding novelty. Moreover, outside the context of novelty and diversity, both Chapelle et al. and Yilmaz at al. [19] demonstrate the value of penalizing redundancy even in the absence of explicit diversity, when there are no subtopics and relevance is assessed against a single overall topic.

Following these suggestions, we deconstruct and analyze proposed measures of novelty and diversity by first considering diversity in Section 2.1 and then novelty in Section 2.2. In Section 2.3 we illustrate a key theoretical shortcoming of the intent-aware versions of more traditional measures, including MAP-IA, when compared to cascade measures. In Section 2.4 we examine the normalization required to meaningfully average these measures across multiple queries.

2.1 Measuring Diversity

Assume a given query has M subtopics. Clarke et al. [7,9], Agrawal et al. [1], and Chapelle et al. [5] all model diversity by assigning a probability p_i, $1 \leq i \leq M$, to each subtopic, indicating the probability that a user entering the query is seeking information related to subtopic i.

Agrawal et al. [1] consider the case when queries are strictly ambiguous and a user is never interested in more than one interpretation of the query. In this case, $\sum_{i=1}^{M} p_i = 1$. They propose a family of measures, each based on a traditional measure such as nDCG [11], average precision, or precision@N. The traditional measure is applied to each subtopic independently and the results are combined to give the expected value of the measure across all users. Thus, we may express the overall score $\mathcal{S}$ for a query as

$$\mathcal{S} = \sum_{i=1}^{M} p_i \, \mathcal{S}_i, \tag{1}$$

where $\mathcal{S}_i$ represents the value of a traditional measure computed over subtopic i. Agrawal et al. call these measures *intent-aware* (IA) versions of the traditional measures. For example, intent-aware average precision ("MAP-IA") would be defined as

$$\text{MAP-IA} = \sum_{i=1}^{M} p_i \, \mathcal{S}_i^{(\text{MAP})}, \tag{2}$$

where $\mathcal{S}_i^{(\text{MAP})}$ represents the value of average precision[2] computed over subtopic i. Chapelle et al. [5] follow the approach of Agrawal et al., but replace $\mathcal{S}_i$ with their own *expected reciprocal rank* (ERR) measure, which penalizes redundancy.

Clarke et al. [7] examine the case when queries are underspecified, and subtopics represent different facets of a single overall interpretation. They assume that a user's interest in one subtopic is independent of her interest in others, so that $\sum_{i=1}^{M} p_i$ may be greater than 1. Nonetheless, they justify a weighted linear combination of novelty scores computed over individual subtopics, which they call α-nDCG

$$\alpha\text{-nDCG} = \frac{\sum_{i=1}^{M} p_i \, \mathcal{S}_i^{(\alpha\text{-nDCG})}}{\mathcal{N}^{(\alpha\text{-nDCG})}}. \tag{3}$$

$\mathcal{N}^{(\alpha\text{-nDCG})}$ is a normalization factor, and $\mathcal{S}_i^{(\alpha\text{-nDCG})}$ is a measure of novelty with respect to subtopic i that penalizes redundancy in a manner similar to that of Chapelle et al. The normalization factor is intended to map the score into the range [0:1], and is related to the normalization used for nDCG [11]. Clarke et al. [9] extend this approach by considering both ambiguous and underspecified (faceted) queries. Inspired by the *rank-biased precision* (RBP) measure of Moffat and Zobel [12], they modify both the novelty measure of α-nDCG and its normalization factor to give a measure they call *novelty- and rank-biased precision* (NRBP).

All of these measures can be cast into a common framework, into which we may substitute competing approaches to novelty and normalization:

$$\mathcal{S} = \frac{\sum_{i=1}^{M} p_i \, \mathcal{S}_i}{\mathcal{N}}. \tag{4}$$

$\mathcal{S}_i$ represents a novelty measure and $\mathcal{N}$ represents a normalization factor. Note that for the TREC 2009 Web Track topics, the p_i are assumed to be equal and may be dropped from the formula.

2.2 Measuring Novelty

Assume we are measuring retrieval effectiveness for a ranked list of K documents $d_1 d_2 ... d_K$. Clarke et al. [7,9] and Chapelle

[2] Unlike the standard acronyms for other effectiveness measures we discuss in this paper, the acronym MAP for *mean average precision* explicitly indicates that the (arithmetic) mean is computed across multiple topics. For simplicity and consistency with other measures, we use this acronym even when considering the average precision over a single topic.

et al. [5] measure novelty indirectly, by penalizing redundancy. On the other hand, as we illustrate in Section 2.3, the intent-aware measures of Agrawal et al. [1] do not measure novelty because they are built upon traditional effectiveness measures.

To penalize redundancy, Chapelle et al. imagine users reading a result list in order, stopping when they find the information they seek. Let q_i^k be the probability that a user who is interested in subtopic i will be satisfied with document d_k. The probability that a user interested in subtopic i will stop at document k is thus

$$\mathcal{Q}_i^k = q_i^k \prod_{j=1}^{k-1} (1 - q_i^j). \tag{5}$$

Clarke et al. [7, 9] arrive at the same formula following a slightly different line of reasoning.

We may treat $\mathcal{Q}_i^k$ as a gain value for the document at rank k. Higher probabilities for documents at higher ranks penalize redundancy by lowering this gain value at lower ranks. The gain value may be discounted by a factor $\mathcal{D}_k$ that depends on rank, accounting for the extra effort required to scan lower ranks and for the possibility that the user will abandon the query without finding anything satisfactory. This discounted gain value is thus $\mathcal{G}_i^k = \mathcal{Q}_i^k / \mathcal{D}_k$. The effectiveness score for the top K documents may then be computed by summing the gain values for each document

$$S_i \; = \; \sum_{k=1}^{K} \mathcal{G}_i^k \; = \; \sum_{k=1}^{K} \frac{\mathcal{Q}_i^k}{\mathcal{D}_k}. \tag{6}$$

In the TREC 2009 Web Track, documents were judged to depth $K = 20$, and we compute measures to this depth in the experimental comparison of Section 3.

Values for q_i^k may be estimated from editorial relevance assessments. Let g_i^k be the editorial relevance grade for document k with respect to subtopic i. Chapelle et al. assume graded relevance assessments are available and define $q_i^k = R(g_i^k)$, where

$$R(g) \; = \; \frac{2^g - 1}{2^G}, \quad g \in \{0, ..., G\}. \tag{7}$$

Clarke et al. assume binary relevance assessments and define

$$q_i^k \; = \; \alpha g_i^k, \quad g \in \{0, 1\}, \tag{8}$$

where $0 < \alpha \leq 1$ is a constant. The TREC 2009 Web Track followed this second approach, using binary assessments and choosing a default value of $\alpha = 0.5$. Thus, if document k is judged relevant to subtopic i, we assume the user agrees only half the time. In the experimental comparison, we explore the impact of varying α.

To combine formula 5 and 8, we first define $c_j^k \; = \; \sum_{j=1}^{k-1} g_i^j$ as the number of documents ranked before position k that are judged relevant to subtopic i. The gain value $\mathcal{Q}_i^k$ may then be computed as

$$\mathcal{Q}_i^k = q_i^k \prod_{j=1}^{k-1} (1 - q_i^j) = \alpha g_i^k \prod_{j=1}^{k-1} (1 - \alpha g_i^k) = \alpha g_i^k (1 - \alpha)^{c_j^k} \tag{9}$$

We use this formula to compute gain values throughout the remainder of the paper.

In the above discussion, we view α as a probability that a user would be satisfied with a judged relevant document. In terms of a user model, we might alternatively view the value $1 - \alpha$ as representing the user's tolerance for redundancy. As α is decreased, the user becomes more willing to accept documents about previously seen subtopics.

Various values for the discount $\mathcal{D}_k$ have been proposed in prior work. Inspired by the discount function employed in the nDCG measure proposed by Järvelin and Kekäläinen [11], Clarke et al. [7] suggest a logarithmic discount of $\mathcal{D}_k = \log_2(k + 1)$. Chapelle et al. [5] suggest a linear discount of $\mathcal{D}_k = k$, producing a reciprocal rank reduction in gain values. Inspired by the rank-biased precision measure proposed by Moffat and Zobel [12], Clarke et al. [9] suggest an exponential discount of $\mathcal{D}_k = (1/\beta)^{k-1}$, where $0 \leq \beta \leq 1$.

From the perspective of a user model, the ratio $\mathcal{D}_k / \mathcal{D}_{k+1}$ represents the probability that a user examining the document at rank k will continue on to examine the document at rank $k + 1$. For the exponential discount, this probability is constant at all ranks $\mathcal{D}_k / \mathcal{D}_{k+1} = \beta^k / \beta^{k-1} = \beta$. For the logarithm and linear discounts, the probability increases at deeper ranks.

2.3 Comparison with Non-Cascade Measures

We may now, by way of example, illustrate the key theoretical difference between the cascade measures and intent-aware versions of more traditional measures, including MAP-IA. As stated previously, the cascade measures reward novelty by penalizing redundancy, as seen in Equations 5 and 9. For our example below, we use Equation 9 with $\alpha = 50\%$.

Consider an ambiguous query with two interpretations ($M = 2$). Assume users entering this query want the first interpretation with probability $p_1 = 60\%$ and the second interpretation with probability $p_2 = 40\%$. Suppose we have a "perfect" document collection. Since the query is ambiguous, no document in this perfect collection can be relevant to both interpretations, but otherwise we assume the collection contains an unlimited number of documents relevant to either of the individual interpretations. In what order should we present these documents to maximize the value of various effectiveness measures?

Unsurprisingly, to maximize any of the measures we examine in this paper, the top-ranked document should be relevant to the most popular interpretation. Under the assumptions of the cascade model, $\alpha = 50\%$ of users who are interested in this first interpretation, or $\alpha p_1 = 30\%$ of all users, will be satisfied by this document. The information needs of the other $(1 - \alpha)p_1 + p_2 = 70\%$ of the users remain unsatisfied, but the proportion of unsatisfied users interested in each interpretation has shifted. Of these unsatisfied users, $(1 - \alpha)p_1/((1 - \alpha)p_1 + p_2) = 43\%$ are now interested in the first interpretation and $p_2/((1 - \alpha)p_1 + p_2) = 57\%$ are now interested in the second interpretation. Thus, to maximize any of the cascade measures, the second-ranked document should be relevant to the second interpretation.

Following this line of reasoning for the deeper ranks, produces a ranked list in which the interpretations are interleaved (1,2,1,2,...). While the most popular interpretation is given the top rank, both interpretations receive good coverage. On the other hand, under the assumptions of intent-aware versions of traditional measures, including MAP-IA, the gain associated with a given document does not reflect the relevance of other documents appearing with it in the result list. For these measures, the value of the measure may be maximized by returning a sequence of documents relevant to the first interpretation (1,1,1,...). Although 40% of users are interested in the second interpretation, their needs would never be satisfied. In general, to obtain a maximum score under the cascade measures even the most unpopular subtopic will eventually appear in the result list; under MAP-IA only the most popular subtopic would ever appear.

2.4 Normalization

To this point, we have considered novelty and diversity with respect to a single query. If an evaluation involved only a single query, no normalization would be required. We could simply set $\mathcal{N} = 1$ and compare runs using their raw scores over the single query. The run with the highest score has the best performance.

When an experiment involves multiple queries, the raw scores $\sum_{i=1}^{M} p_i \, \mathcal{S}_i$ must be normalized into the range [0:1] for the scores to be meaningfully averaged. This normalization is required not only for conventional reasons but also because of the influence that the number of subtopics, and their associated weights, can have on the value of the raw scores. We illustrate this influence with an example based on the NRBP measure. For the purpose of this example, assume we have a query where $p_i = p$, a constant for all subtopics, $1 \leq i \leq M$. Remember that for a faceted query, p need not be $1/M$, since $\sum_{i=1}^{M} p_i$ may be greater than 1.

As we did in the last section, imagine we are searching a perfect collection containing an unlimited number of documents relevant to all subtopics. The highest possible score an ideal result list could achieve over this perfect collection may be computed as:

$$\begin{aligned} \textit{raw score} &= \sum_{i=1}^{M} p_i \sum_{k=1}^{K} \frac{\mathcal{Q}_i^k}{\mathcal{D}_k} \qquad (10) \\ &= p \sum_{i=1}^{M} \sum_{k=1}^{K} \frac{\alpha(1-\alpha)^{k-1}}{(1/\beta)^{k-1}} \\ &= \alpha p M \sum_{k=0}^{K-1} (\beta(1-\alpha))^k \\ &\rightarrow \frac{\alpha p M}{1-(1-\alpha)\beta}, \text{ as } K \rightarrow \infty. \end{aligned}$$

The value of the highest possible score grows linearly with both M and p, suggesting that we might expect larger raw scores from queries with more subtopics and larger weights. Other cascade measures exhibit similar properties.

For normalization, raw scores are divided by the highest possible score achievable. Clarke et al. [9] discuss two types of normalization: 1) *collection-dependent* normalization, which they call “ideal” normalization, and 2) *collection-independent normalization*, which they call “ideal ideal” normalization. Collection-dependent normalization is based on the highest possible score achievable from known relevant documents in a specific test collection. Collection-independent normalization is based on the highest possible score achievable from a perfect collection.

For example, for topic 47 in Figure 1 there are 18 documents that are known to be relevant to both subtopic 1 and subtopic 2 in the ClueWeb09 collection. A further 114 documents are relevant only to subtopic 1, and 33 documents are relevant only to subtopic 2. No documents are known to be relevant to subtopic 3. The ideal collection-dependent result would thus have the first 18 ranks filled with the documents relevant to both subtopics 1 and 2. The next 66 ranks would alternate between the subtopics, which would then be followed by the remaining 48 documents relevant only to subtopic 1. On the other hand, an ideal collection-independent result would assume the existence of an unlimited number of documents relevant to all three subtopics.

For the experimental work reported in this paper, we compute collection-independent normalization by assuming the existence a perfect collection containing an unlimited number of documents relevant to all subtopics. We adhere to this assumption even for ambiguous queries such as topic 20 in Figure 1, and even for navigational subtopics, such as subtopic 1 of topic 20, where only a single relevant document could exist. While it is certainly possible for collection-independent normalization to take the characteristics of topics and subtopics into account, handling ambiguous queries differently than faceted queries and navigational subtopics differently than informational subtopics, we leave the exploration of this idea to future work.

Theoretically, collection-independent normalization provides benefits over collection-dependent normalization. As shown by a number of authors, the computation of collection-dependent normalization is NP-hard in the general case [4, 6, 7, 21], although a simple greedy approximation is usually acceptable in practice. The value of a collection-dependent normalization factor may change substantially if new relevant documents are discovered in the collection; collection-independent normalization may be computed without any knowledge of documents outside the result list. Nonetheless, collection-dependent normalization more accurately reflects the reality of a specific collection, which may diverge considerably from the assumptions of a perfect collection. Moreover, Carterette [4] demonstrates that even the greedy approximation of collection-dependent normalization can substantially impact the value of novelty and diversity measures, at least for some topics.

As is typical for IR evaluation measures, we average across topics using an arithmetic mean. However, to compute MAP Robertson [15] advocates for the geometric mean, demonstrating that it tends to emphasize performance on harder topics. In light of our concerns with normalization, the application of the geometric mean also might be appropriate for averaging cascade measures. Since the geometric mean is computed by multiplying scores across topics, the normalization factors can be ignored. Of course, scores of zero must be appropriately handled, and we leave the exploration of this idea to future work.

2.5 Summary of Measures

Figure 2 provides a summary of intent-aware measures examined in the experimental comparison of Section 3. The general form of the measures appears on the far right. For the cascade measures, the upper portion of the diagram indicates how each measure is constructed from its components, including discounting and normalization (collection dependent or independent). For MAP-IA, $\mathcal{S}_i$ is computed as the traditional MAP value for subtopic i. Since MAP already incorporates a collection-dependent normalization — taking into account the number of known relevant documents — we normalize MAP-IA by averaging across subtopics.

Note that the measure we call ERR differs slightly from the intent-aware version of the measure proposed by Chapelle et al. [5]. Since graded relevance values are not available, our version estimates gain values through Equation 9 instead of

diversity	novelty	gain	discount	normalization ($\mathcal{N}$)	measure
$\mathcal{S} = \frac{\sum_{i=1}^{M} p_i \, \mathcal{S}_i}{\mathcal{N}}$	$\mathcal{S}_i = \sum_{k=1}^{K} \frac{\mathcal{Q}_i^k}{\mathcal{D}_k}$	$\mathcal{Q}_i^k = q_i^k \prod_{j=1}^{k-1}(1 - q_i^j)$ simplified to $\mathcal{Q}_i^k = \alpha g_i^k (1-\alpha)^{c_j^k}$	$\mathcal{D}_k = \log(k+1)$	independent	α-DCG
				dependent	α-nDCG
			$\mathcal{D}_k = k$	independent	ERR
				dependent	nERR
			$\mathcal{D}_k = (1/\beta)^{k-1}$	independent	NRBP
				dependent	nNRBP
	$\mathcal{S}_i$ = traditional MAP over subtopic i			$\mathcal{N} = M$	MAP-IA

Figure 2: Summary of intent-aware measures examined in our experimental comparison.

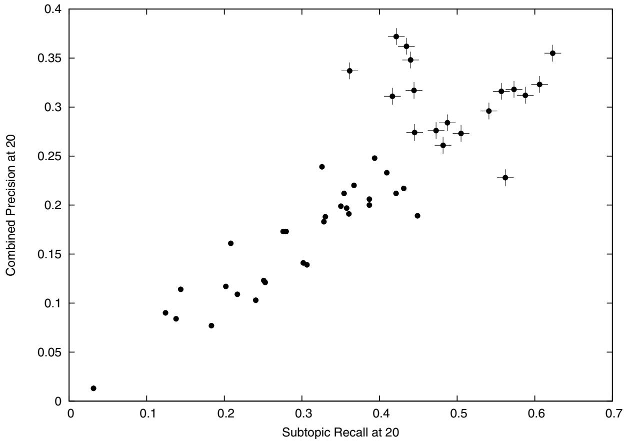

Figure 3: Comparison of subtopic recall and combined precision@20 for 49 runs submitted to the diversity task of the TREC 2009 Web Track. Each point represents the average performance over 50 topics for a run submitted to the TREC 2009 Web track. The points overstruck with a plus sign ("+") will be the focus of the experiments reported in the remainder of this paper.

Equation 7. In addition, our version incorporates normalization for the reasons outlined in Section 2.4.

For our experimental comparison, all measures are computed to retrieval depth $K = 20$. A default value of $\alpha = 0.50$ is used unless otherwise indicated. For NRBP and nNRBP a value of $\beta = 0.8$ (a relatively patient user) is adopted for all experiments.

3. EXPERIMENTAL COMPARISON

Figure 3 presents the 49 runs submitted for the diversity task of the TREC 2009 Web Track [8]. The figure compares the runs when evaluated by two simple measures — mean *subtopic recall* and mean *combined precision* — with both measures computed at a depth of $K = 20$ documents. To compute subtopic recall for a given topic, we count the number of subtopics with a relevant document in the top K and divide by the total number of subtopics for that topic. To compute combined precision for a given topic, we count the number of documents relevant to any subtopic in the top K and divide by K. The computation of combined precision ignores the distinction between subtopics and treats a document as relevant to the topic as a whole if it is relevant to any subtopic.

As we see in Figure 3, these simple effectiveness measures are generally correlated, particularly at the lower end of the performance range. However, at the upper end of the performance range the correlation appears much weaker. Consider the 18 points overstruck with a plus sign ("+") in the upper right of the plot. These 18 runs all provide solid performance in terms of one or both measures. Nonetheless, the best run in terms of combined precision ranks 19th in terms of subtopic recall, and the fifth-best run in terms of subtopic recall ranks 21st in terms of combined precision. We expect that these differences would be noticeable and meaningful to a user. Within the 18 selected runs, the subtopic recall of the top run represents a 72% improvement over the bottom run, and the combined precision of the top run represents a 63% improvement over the bottom run.

These two simple measures provide two contrasting views on effectiveness. In some sense, the purpose of a novelty and diversity measure is to combine these views in a sensible way. Given their reasonable performance on at least one of our simple measures, we focus our attention on the 18 points in the upper left.

3.1 Comparison Between Measures

We first compare the novelty and diversity measures presented in Section 2 to the simple measures discussed above. We use Kendall's τ to measure the stability of the rankings of experimental runs under different effectiveness measures. Kendall's τ is a well-established rank correlation measure for comparing such rankings [2, 17, 18]. Values range from +1 to −1, with +1 indicating perfect agreement and −1 indicating the opposite. Given a pair of rankings, prior work views a τ value of 0.9 or higher as indicating that the rankings are "equivalent" and values below 0.8 as indicating that the rankings contain "noticeable differences" [2].

Each graph in Figure 4 plots the Kendall τ correlation between one of the cascade measures and the two simple measures discussed above. For this comparison we use collection-dependent normalization. Note that the value of α can range as low as 0 because a factor of α cancels from the gain values during normalization. As the value of α increases, the correlation with combined precision drops and the correlation with subtopic recall grows. This behavior is consistent with the idea that $1 - \alpha$ represents the user's tolerance for redundancy. We may view combined precision as modeling an extremely tolerant user, who never penalizes redundancy, while subtopic retrieval models an extremely intolerant user,

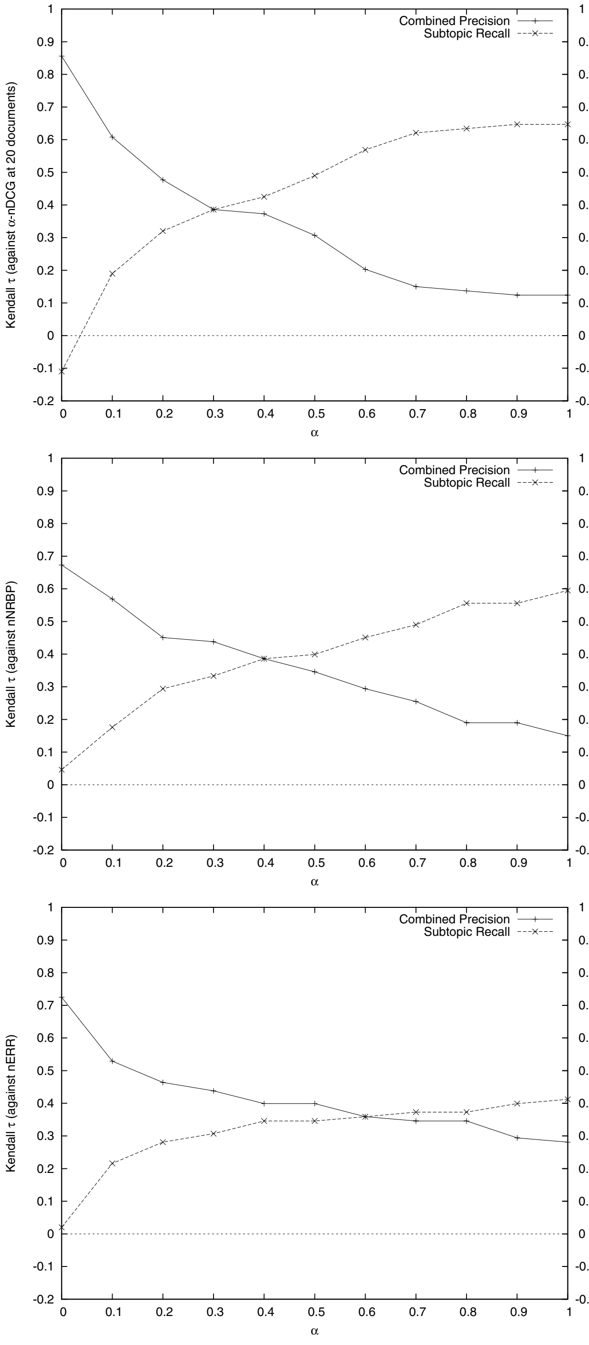

Figure 4: The impact of varying α. Each plot varies α for one of the collection-dependent cascade measures presented in Figure 2. As α increases from 0 to 1, the Kendall τ correlation (x axis) with subtopic recall increases and the correlation with overall precision decreases. The value of α can range as low as 0 because a factor of α cancels from the gain values during normalization.

who never accepts redundancy. Intermediate values of α appear to balance the two extremes. The stronger correlations seen with α-nDCG appear to be related to the weaker discount used by that measure. When the measures are computed at shallower depths ($K < 20$) the Kendall τ values change (not shown) but the trends are the same.

Figure 5 shows the correlation between the main novelty and diversity measures. The three cascade measures are highly correlated with each other, but not with MAP-IA. Figure 6 shows the impact of normalization. Over this test collection, we see little difference between collection-dependent and collection-independent normalization, which give rankings that are essentially equivalent.

3.2 Discriminative Power

Confirming and extending work by Sakai et al. [17] we examine the *discriminative power* of the various measures, giving additional consideration to comparisons with traditional measures. Sakai [16] proposes a simple method for assessing the discriminative power of effectiveness measures. The method computes a significance test between every pair of experimental runs and reports the percentage of pairs that are significant at some fixed significance level. For the experiments in this paper, we fix the significance level at 0.05.

The results are presented in Figure 7. We apply both the two-tailed paired t-test and bootstrap as our significance tests. The results show, for example, that 30.7% of the pairs would have a significant difference under α-nDCG and the two-tailed paired t-test. Sakai's measure of discriminative power suggests the degree to which a given effectiveness measure can detect differences between runs relative to other measures. Sakai et al [17] consider "high discriminative power as a necessary condition for a good evaluation metric, not as a sufficient condition." Given the consistency of the results between the bootstrap and the t-test, we focus the remainder of our discussion on the bootstrap results.

Recall that the 18 runs selected for our experimental study all produce solid performance on one or both or our simple effectiveness measures: combined precision and subtopic recall. For these runs, the discriminative power of combined precision is much lower than that of subtopic recall. However, the figure also reports traditional MAP computed using the same approach that we used for combined precision, by ignoring the distinction between subtopics and treating a document as relevant to the topic as a whole if it is relevant to any subtopic.

Both traditional MAP and subtopic recall have discriminative power above 40%, while all of the cascade measures have discriminative power below 33% and MAP-IA has discriminative power below 35%. These results were unexpected since the cascade measures and MAP-IA incorporate information individually unavailable to MAP and subtopic recall. MAP is computed without considering any distinction between subtopics; subtopic recall is computed without considering multiple relevant documents for subtopics. In light of these results, we suggest that future evaluation efforts continue to report a variety of measures, including simple measures and more traditional measures.

4. CONCLUDING DISCUSSION

Section 2 places cascade measures of novelty and diversity into a unified framework, founded upon simple models of user behavior. Queries are decomposed into subtopics, Correlations between the major novelty and diversity measures presented in Section 2 and Figure 2. Correlations between the three cascade measures are quite high, with Kendall τ values ranging from 0.882 to 0.892, close to equivalent. Correlations between the cascade measures and MAP-IA are much lower, with Kendall τ values ranging from 0.216 to 0.307.

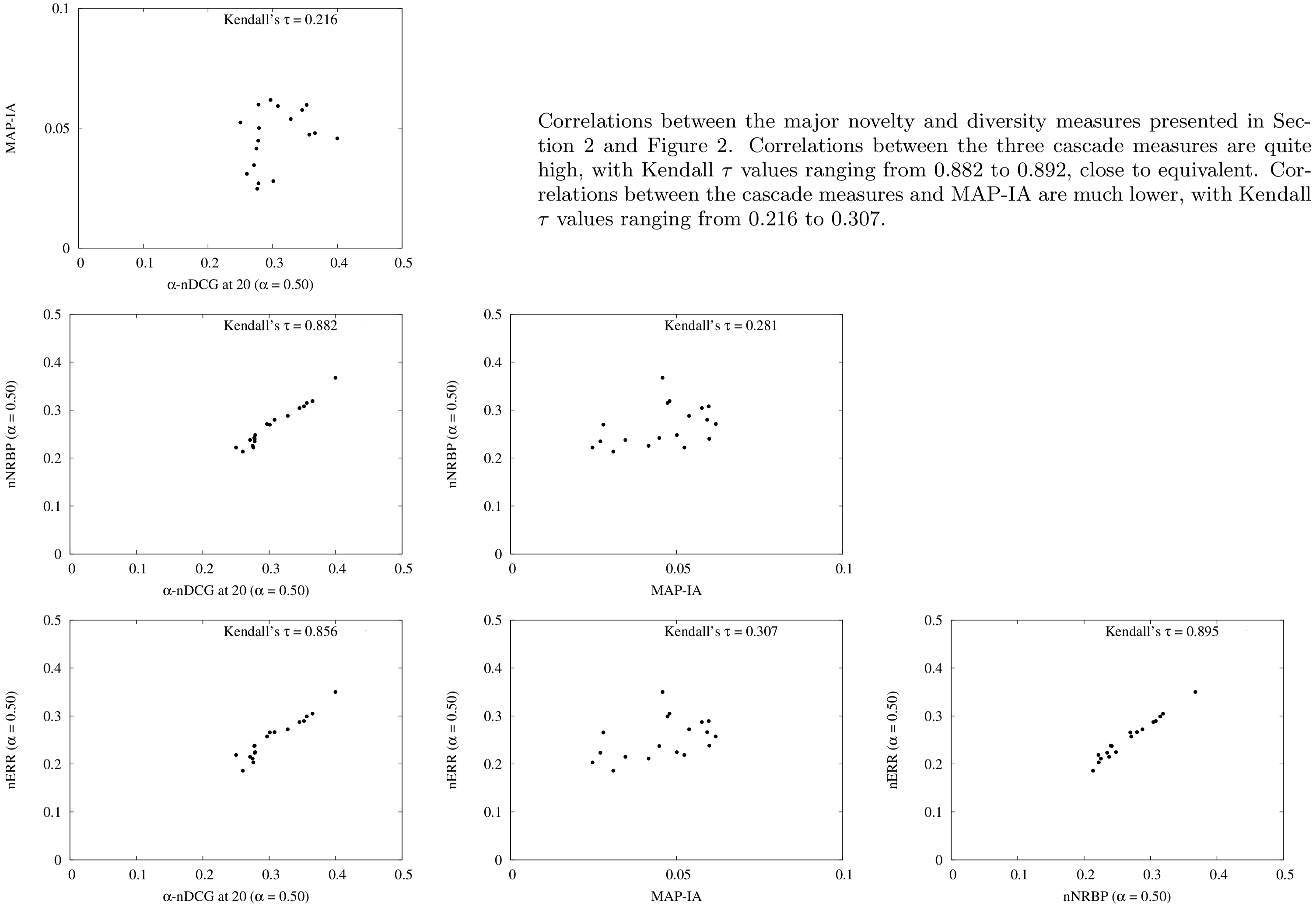

Figure 5: Correlations between novelty and diversity measures.

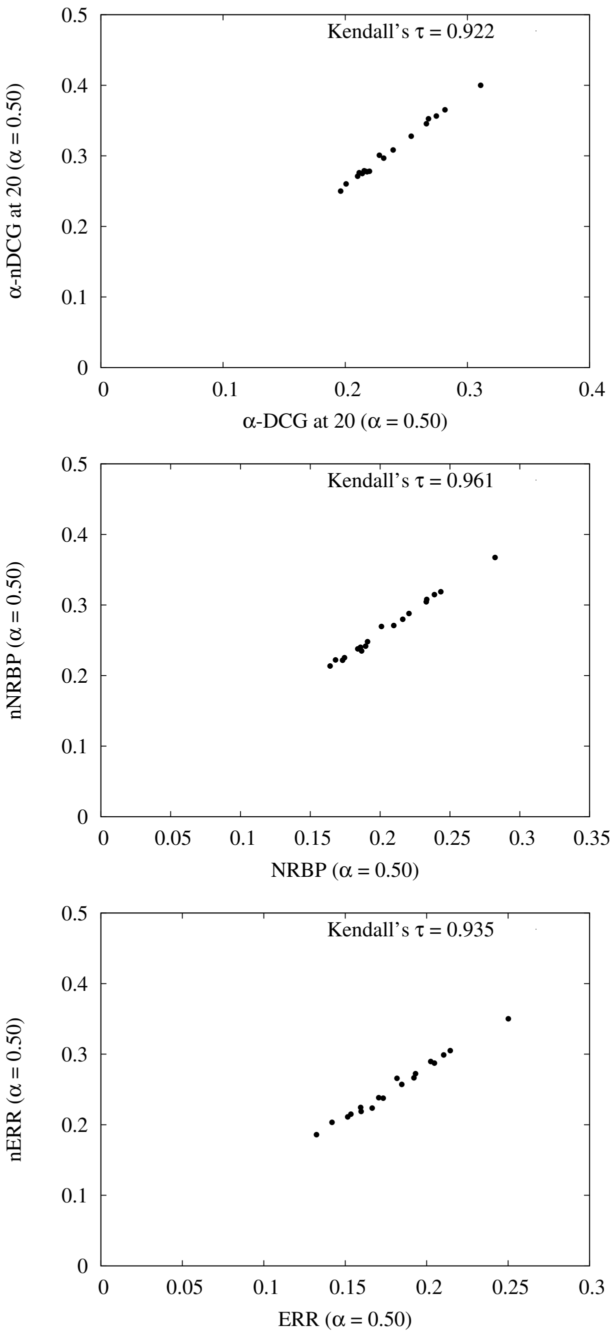

Figure 6: The impact of normalization. Each plot compares collection-independent normalization (x axis) against collection-dependent normalization (y axis) for one of the cascade measures presented in Section 2. The correlation between collection-independent normalization and collection-dependent normalization is high, with Kendall τ values above 0.922**, essentially equivalent.**

	bootstrap	**paired t-test**
Combined precision	20.9%	22.2%
Subtopic recall	44.4%	46.4%
α-DCG	30.7%	32.0%
α-nDCG	32.7%	34.0%
ERR	28.8%	29.4%
nERR	27.5%	28.8%
NRBP	28.7%	29.4%
nNRBP	30.1%	30.7%
MAP-IA	34.6%	37.9%
MAP	41.8%	49.0%

Figure 7: Discriminative power of measures under the two-tailed paired t-test and bootstrap tests with a significance level of 0.05.

each with an associated probability p_i that a user entering the query is seeking information related to subtopic i. Diversity is accommodated through a linear combination of measures computed on individual subtopics, weighted according to these probabilities.

Novelty is accommodated by penalizing redundancy. Given a ranked list of documents, the document at rank k has an associated probability q_i^k that a user who is interested in subtopic i will be satisfied by that document. Assuming that a satisfied user is no longer interested in documents about that subtopic, we may compute a gain value at each rank, according to these probabilities. At rank k, the gain is discounted by $\mathcal{D}_k$ to reflect the probability that a user might stop before examining the document. We demonstrate the importance of penalizing redundancy when computing gain values by highlighting limitations in competing approaches.

Since different queries may have different numbers of subtopics, weighted according to different probabilities, we show that normalization is required to sensibly average measures across multiple queries. Following the ideas of Järvelin and Kekäläinen [11], we normalize against an ideal gain vector, which may be collection dependent or collection independent. The collection-dependent gain vector is computed from known relevant documents; the collection-independent gain vector is computed by imagining a perfect collection containing an unlimited number of documents relevant to all subtopics. As future work, we plan to further explore the need for collection-dependent normalization, examine alternative methods for computing collection-independent normalization, and evaluate the use of the geometric mean for averaging across topics.

The test collection and runs created through the TREC 2009 Web track provide a vehicle for exploring and validating these cascade measures. Our experimental comparison indicates that these measures work as intended, rewarding systems that achieve a balance between subtopic recall and combined precision. Kendall τ correlations between the cascade measures are high, and we have found no evidence favoring one over another. High correlations between collection-dependent and collection-independent normalizations suggest that collection-dependent normalization may not be required. Unfortunately, our results indicate that the discriminative power of the cascade measures may be lower than that measures such as traditional MAP and subtopic recall, suggesting that future evaluation efforts continue to report a variety of measures. The TREC Web Track continues in 2010, with ERR as its primary evaluation measure.

As future work we plan to explore new methods for computing gain and discount values and seek additional validation of the existing methods. Outside the context of novelty and diversity, click logs and other information from commercial search engines have been applied to validate effectiveness measures that incorporate simple user models [5,19]. Zhang et al. [23] use click logs to validate discount functions. In addition, we plan to compare the cascade measures to other simple methods for evaluating novelty and diversity, including ideas by Zhai et al. [21], Chen and Karger [6], and Sakai et al. [17].

5. ACKNOWLEDGMENTS

We thank the TREC Web track participants for their enthusiastic acceptance of our efforts to create and validate a methodology for evaluating novelty and diversity. We received many comments from Web track participants and others. We especially thank Olivier Chapelle, Don Metzler, Evangelos Kanoulas, Mark Smucker, Gord Cormack, and Maheedhar Kolla.

6. REFERENCES

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *2nd ACM International Conference on Web Search and Data Mining*, pages 5–14, 2009.

[2] C. Buckley and E. M. Voorhees. Retrieval evaluation with incomplete information. In *27th Annual International ACM SIGIR Conference*, pages 25–32, 2004.

[3] J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *21st Annual International ACM SIGIR Conference*, pages 335–336, 1998.

[4] B. Carterette. An analysis of NP-completeness in novelty and diversity ranking. In *2nd International Conference on the Theory of Information Retrieval*, pages 200–211, 2009.

[5] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *18th ACM Conference on Information and Knowledge Management*, pages 621–630, 2009.

[6] H. Chen and D. R. Karger. Less is more: Probabilistic models for retrieving fewer relevant documents. In *29th Annual International ACM SIGIR Conference*, pages 429–436, 2006.

[7] C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkann, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *31st Annual International ACM SIGIR Conference*, pages 659–666, 2008.

[8] C. L. A. Clarke, N. Craswell, and I. Soboroff. Overview of the TREC 2009 Web track. In *18th Text REtrieval Conference*, 2009.

[9] C. L. A. Clarke, M. Kolla, and O. Vechtomova. An effectiveness measure for ambiguous and underspecified queries. In *2nd International Conference on the Theory of Information Retrieval*, pages 188–199, 2009.

[10] N. Craswell, O. Zoeter, M. J. Taylor, and B. Ramsey. An experimental comparison of click position-bias models. In *International Conference on Web Search and Web Data Mining*, pages 87–94, 2008.

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

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

[13] F. Radlinski, R. Kleinberg, and T. Joachims. Learning diverse rankings with multi-armed bandits. In *25th International Conference on Machine Learning*, pages 784–791, 2008.

[14] D. Rafiei, K. Bharat, and A. Shukla. Diversifying Web search results. In *19th International World Wide Web Conference*, 2010.

[15] S. Robertson. On GMAP: and other transformations. In *15th ACM International Conference on Information and Knowledge management*, pages 78–83, 2006.

[16] T. Sakai. Evaluating evaluation metrics based on the bootstrap. In *29th Annual International ACM SIGIR Conference*, pages 525–532, 2006.

[17] T. Sakai, N. Craswell, R. Song, S. Robertson, Z. Dou, and C.-Y. Lin. Simple evaluation metrics for diversified search results. In *3rd International Workshop on Evaluating Information Access*, 2010.

[18] E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. In *21st Annual International ACM SIGIR Conference*, pages 315–323, 1998.

[19] E. Yilmaz, M. Shokouhi, N. Craswell, and S. Robertson. Incorporating user behavior information in IR evaluation. In *SIGIR 2009 Workshop on Understanding the User: Logging and Interpreting User Interactions in Information Retrieval*, 2009.

[20] E. Yilmaz, M. Shokouhi, N. Craswell, and S. Robertson. Expected browsing utility for Web search evaluation. In *19th ACM International Conference on Information and Knowledge Management*, 2010.

[21] C. Zhai, W. W. Cohen, and J. Lafferty. Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. In *26th Annual International ACM SIGIR Conference*, pages 10–17, 2003.

[22] B. Zhang, H. Li, Y. Liu, L. Ji, W. Xi, W. Fan, Z. Chen, and W.-Y. Ma. Improving web search results using affinity graph. In *28th Annual International ACM SIGIR Conference*, pages 504–511, 2005.

[23] Y. Zhang, L. A. F. Park, and A. Moffat. Click-based evidence for decaying weight distributions in search effectiveness metrics. *Information Retrieval*, 13(1):46–69, February 2010.

[24] X. Zhu, A. B. Goldberg, J. Van Gael, and D. Andrzejewski. Improving diversity in ranking using absorbing random walks. In *Proceedings of Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 97–104, 2007.