Diversity by Proportionality: An Election-based Approach to Search Result Diversification

Van Dang and W. Bruce Croft
Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts
Amherst, MA 01003
{vdang, croft}@cs.umass.edu

ABSTRACT

This paper presents a different perspective on diversity in search results: diversity by proportionality. We consider a result list most diverse, with respect to some set of topics related to the query, when the number of documents it provides on each topic is proportional to the topic's popularity. Consequently, we propose a framework for optimizing proportionality for search result diversification, which is motivated by the problem of assigning seats to members of competing political parties. Our technique iteratively determines, for each position in the result ranked list, the topic that best maintains the overall proportionality. It then selects the best document on this topic for this position. We demonstrate empirically that our method significantly outperforms the top performing approach in the literature not only on our proposed metric for proportionality, but also on several standard diversity measures. This result indicates that promoting proportionality naturally leads to minimal redundancy, which is a goal of the current diversity approaches.

Categories and Subject Descriptors

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

General Terms

Algorithms, Measurement, Performance, Experimentation.

Keywords

Search result diversification, proportional representation, proportionality, redundancy, novelty, Sainte-Laguë.

1. INTRODUCTION

Search result diversification techniques have been studied as a method of tackling queries with unclear information needs. Standard retrieval models and evaluations are based on the assumption that there is a single specific topic associated with the relevant documents for a query. Diversification models [4, 1, 26], on the other hand, identify the probable "aspects" of the query and return documents for each of these aspects, making the result list more diverse. Aspects denote the multiple possible intents, interpretations, or subtopics associated with a given query. By explicitly representing and providing diversity in the result list, these models can increase the likelihood that users will find documents relevant to their specific intent and thereby improve effectiveness.

This problem of finding a diverse ranked list of documents, with respect to the aspects of the query, has been studied primarily from the perspective of minimizing *redundancy*. In other words, existing work focuses on penalizing result lists with too many documents on the same aspect, which increases the redundancy of coverage, and promoting lists that contain documents covering multiple aspects. Most of the effectiveness measures for diversity [7, 8] are also based on this notion of redundancy. They penalize the redundancy in a ranked list of documents by judging each of the documents given the context of those retrieved at earlier ranks [9].

In this paper, we approach the same task from a different perspective. We view the problem of finding a good result list of any given size as the task of finding a representative sample of the larger set of ranked documents. Hence, the quality of the subset (a result list) should be measured by how well it represents the whole set (a much larger sample of the ranking). Using a simple (and well-worn) example, in a ranked list for a query "java", 90% of the documents may be about the programming language and 10% about the island. From our perspective, a result list containing ten documents where only one of them was about the island would be more representative than a result list containing five documents on each subtopic. Consequently, we treat the problem of finding a diverse result of documents as finding a *proportional representation* for the document ranking.

Finding a *proportional representation* is a critical part of most electoral processes. The problem is to assign a set of seats in the parliament to members of competing political parties in a way that the number of seats each party possesses is proportional to the number of votes it has received. In other words, the members in the elected parliament must be a *proportional representation* of these parties. If we view each position in our ranked list as a "seat", each aspect of the

SIGIR'12, August 12–16, 2012, Portland, Oregon, USA.

query as a “party” and the aspect popularity as the “votes” for this “party”, the problem of diversification becomes very similar to this seat allocation problem.

Based on the above analogy, we propose a novel technique for search result diversification. It is an adaptation of the Sainte-Laguë method, a standard technique for finding proportional representations that is used in the official election in New Zealand[1]. Generally, our technique starts with an empty ranked list of a certain size. It sequentially visits each “seat” in the list and determines for each of them to which aspect it should be allocated in order to maintain proportionality. Then it selects the best document for the selected aspect to occupy this “seat”. In addition, we also present a new effectiveness measure that captures *proportionality* in search results. We demonstrate empirically that our method is more effective than the top performing approach in the diversity literature not only according to the proportionality measure but also using several standard metrics including α-NDCG [7] and NRBP [8] that existing work has been designed to optimize. This indicates that optimizing search results for proportionality naturally leads to minimal redundancy and a diverse, effective result list.

In the next section, we briefly mention some related work. Section 3 presents our approach to proportionality and the effectiveness measure based on it. Section 4 describes in detail our proportionality-driven framework for search result diversification. Section 5 and 6 contains the experimental setup and results, as well as analyses and discussions. Finally, Section 7 concludes.

2. RELATED WORK

The literature of diversification has been concentrating on the notion of *novelty* and *redundancy*. These two notions are considered under the context of user behavior with the assumption that users examine the result lists top down and eventually stop reading. Therefore, a document at any rank providing the same information as those at earlier ranks is considered *redundant*. Likewise, a *novel* document is one that provides information that has not been covered by any of the previous documents. As a result, a ranked list is considered more diverse if it contains less redundancy, or equivalently, more novelty.

This is clearly demonstrated through several standard effectiveness metrics such as α-NDCG [7] and NRBP [8]. They measure the diversity of a ranked list by explicitly rewarding novelty and penalizing redundancy observed at every rank. Similarly, diversification techniques [4, 29, 1, 5, 26] attempt to form a diverse ranked list by repeatedly selecting documents that are different to those previously selected. In other words, they try to accommodate novelty at every position in the list.

In this paper, we present a different perspective on diversity. This view of diversity emphasizes proportionality, which is the property that the number of documents returned for each of the aspects of the query is proportional to the overall popularity of that aspect. Consequently, the framework we derive is driven by this notion of proportionality, thus is different from the existing work.

Several metrics have been proposed to measure the proportionality in the outcome of an electoral process, an excellent summary of which is provided by Gallagher and Lijphart [17, 19]. They can be classified into two broad categories: the first concentrates on the absolute difference between the percentage of seats and the percentage of votes, the second focuses on the the ratio between them. These measures appear mathematically simple but attempt to address complex specific issues of elections that are not always relevant to our context. As a result, we apply Gallagher's Index [17], or the least square index, which is reasonably suited to our problem.

For completeness, we will provide a brief survey of techniques in the current literature of diversification. They can be classified as either implicit or explicit approaches. The former [4, 29] assumes that similar documents cover similar aspects without modelling the actual aspects. They iteratively select documents that are similar to the query but different to the previously selected documents in terms of vocabulary [4] or divergence from one language model to another [29]. More recent work [25, 22] applies the portfolio theory to document ranking, which views diversification as a mean of risk minimization. Explicit approaches [23, 1, 5, 26], on the other hand, explicitly models aspects of a query with a taxonomy [1], top retrieved documents [5] or query reformulations [23, 26] and thus can directly select documents that cover different aspects. Experimentally, explicit approaches have been demonstrated to be superior to implicit approaches [26].

3. PROPORTIONALITY

In this paper, we view the task of diversification as finding a proportional representation for a document ranking. In this section, we will explain the notion of *proportionality* as well as describing an effectiveness measure for it.

3.1 Definition of Proportionality

Let $T = \{t_1, t_2, ..., t_n\}$ indicate a set of aspects for a query q and D denote a large set of documents related to q. Let p_i indicate the popularity of the aspect $t_i \in T$, which is the percentage of documents in D covering this aspect. Additionally, let S be any subset of D. We define S to be *proportional to* D, or a *proportional representation* of D, with respect to T if and only if the number of documents in S that is relevant to each of the aspects $t_i \in T$ is proportional to its popularity p_i.

Let us revisit the example in Section 1, in which there are 90% of documents in D about the “java” programming language and the rest 10% is about an island named “java”. Let $\{x, y\}$ denote any subset of D with x documents about programming and y documents about the island. In this case, $\{9, 1\}$ is proportional and thus is a proportional representation of D. While $\{8, 2\}$ is not proportional, it is more proportional than $\{7, 3\}$.

Let R indicate a ranking of documents in D and S now represent a sub-ranking of R. We define S to be proportional to R if the subset of documents S provides at every rank is a proportional representation of D.

3.2 Effectiveness Measure

This notion of proportionality is, in fact, frequently used in evaluating the outcome of elections in which seats are assigned to members of competing political parties. This problem can be stated as follows. We have a limited number of seats in the parliament and a number of competing parties. Each party has its own members. Through election

[1] http://www.elections.org.nz/voting/mmp/sainte-lague.html

campaigns, each party obtains a number of votes from people around the country. The goal is to assign members of different parties to the seats such that the number of seats each party gets is proportional to the votes it receives.

Several metrics have been proposed to measure such proportionality. Most of them are based on the difference between the percentage of votes each party receives and the percentage of seats it gets. Among those, the least square index (LSq) [17] is one of the standard metrics for measuring dis-proportionality:

$$LSq = \sqrt{\frac{1}{2}\sum_i (v_i - s_i)^2} \simeq \sum_i (v_i - s_i)^2$$

where v_i and s_i are the percentage of votes and the percentage of seats the i-th party received. Let us illustrate this with an example in which we have *ten* seats and *three* competing parties, namely A, B and C. Let us assume both A and B receive 50% of the votes and C gets 0%. Clearly, the proportional assignment which provides A and B each with *five* seats and C with *none* will result in $LSq = 0$. The value for LSq will increase when the seat assignment becomes more disproportional.

We will now turn our attention to the proportionality of a retrieved list of *ten* documents for the query "satellite", which we assume to have two aspects: "satellite internet" and "satelilte phone" with equal popularity of 50%. Due to the possible presence of non-relevant documents, we have to create a third "aspect" to account for non-relevant documents. As a result, proportionality requires this list to contain *five* relevant documents for each of the two aspects and *zero* documents for the "non-relevant" aspect. This situation seems to be very similar to the election described above. Unfortunately, we cannot apply LSq to measure the dis-proportionality of this result list due to two differences.

First, each member typically belongs to exactly one political party. As a result, one party gets more seats than it should always indicates that some other party is getting less than they deserve. A document, however, might be related to multiple aspects of a query. It then is possible that an aspect can be "rewarded" with additional documents while others still have as many relevant documents as they deserve.

Second, it is equally bad for any party to get any more seats than it should. In our case, however, selecting for the result list a document that is relevant to an aspect that already has enough relevant documents in the list is not as bad as selecting a non-relevant document.

Taking both differences into consideration, we argue that LSq, since is designed for the seat allocation problem, puts too much penalty on overly representing query aspects. LSq fails to recognize that some of these situations do not create any undesirable consequences in our setting, and thus should not be penalized. Therefore, we propose a new metric, disproportionality at rank K, calculated as follows:

$$DP@K = \sum_{aspect\ t_i} c_i (v_i - s_i)^2 + \frac{1}{2} n_{NR}^2 \qquad (1)$$

where v_i is the number of relevant documents that the aspect t_i should have, s_i is the number of relevant documents the system actually found for this aspect, n_{NR} is the number of non-relevant documents, and

$$c_i = \begin{cases} 1 & v_i \geq s_i \\ 0 & \text{otherwise} \end{cases}$$

Formula (1) has two important properties. The first is that it penalizes a result set for under-representing any aspect of the query ($s_i < v_i$) but not for over-representing them ($s_i > v_i$), which addresses the first issue associated with LSq. The second is that while the over-representation of a query aspect is not penalized, the over-representation of the "non-relevant" aspect ($n_{NR} > 0$) is, which overcomes the second issue associated with LSq.

A perfectly disproportional set of documents in the context of information retrieval would be a set with all non-relevant documents. Thus, the Ideal-DP is given by:

$$Ideal\_DP@K = \sum_{aspect\ t_i} v_i^2 + \frac{1}{2} K^2$$

The last step is to derive our proportionality measure by normalizing the DP score with Ideal-DP in order to make it comparable across queries:

$$PR@K = 1 - \frac{DP@K}{IDeal\_DP@K}$$

Finally, the Cumulative Proportionality (CPR) measure for rankings is calculated as follows:

$$CPR@K = \frac{1}{K} \sum_{i=1}^{K} PR@i$$

4. PROPOSED METHOD

In this section, we first introduce the Sainte-Laguë method, a standard technique for finding proportional representations that is used to solve the seat allocation problem described in Section 3.2. We then demonstrate the analogy between this problem and our problem of propotionality-based diversification, which helps us derive our technique from the Sainte-Laguë method.

4.1 The Sainte-Laguë Method

This method considers all of the available seats iteratively. For each of them, it computes a *quotient* for all of the parties based on the votes they receive and the number of seats they have taken. This seat is then assigned to the party with the largest quotient, which helps maintain the overall proportionality. We assume the selected party will then assign one of its members to this seat. Finally, it increases the number of seats assigned to the chosen party by one. The process repeats until all seats are assigned. Pseudo code for this procedure is provided as Algorithm 1. In this procedure, $P = \{P_1, P_2, ..., P_n\}$ is the set of parties and $M_i = \{m_1^{(i)}, m_2^{(i)}, ..., m_{l_i}^{(i)}\}$ is the set of members of the party P_i. v_i and s_i indicate the number of votes P_i receives and the number of seats that have been assigned to P_i so far.

Algorithm 1 The Sainte-Laguë method for seat allocation

```
$s_i \leftarrow 0, \forall i$
for all available seats in the parliament do
    for all parties $P_i$ do
        $quotient[i] = \frac{v_i}{2s_i+1}$
    end for
    $k \leftarrow \arg\max_i quotient[i]$
    $m^* \leftarrow$ the best member of $P_k$
    Assign the current seat to $m^*$
    $M_k \leftarrow M_k \setminus \{m^*\}$
    $s_k \leftarrow s_k + 1$
end for
```

4.2 Diversity by Proportionality

4.2.1 Framework

Let q indicate the user query, $T = \{t_1, t_2, ..., t_n\}$ indicate the aspects for q whose popularity is $\{p_1, p_2, ..., p_n\}$. In addition, let $R = \{d_1, d_2, ..., d_m\}$ be the ranked list of documents returned by an initial retrieval and $P(d_i|t_j)$ indicate some estimate of the probability that the document d_i is relevant to the aspect t_j. The task is to select a subset of R to form a diverse ranked list S of size k.

As mentioned earlier, existing techniques [1, 26] generally favor an S with smaller redundancy. Our idea, on the other hand, is to favor an S with higher proportionality. The optimal S, consequently, is a ranked list in which the number of relevant documents for each of the aspects t_i is proportional to its popularity p_i. This objective is, in fact, very similar to that of the seat allocation problem above. As a result, we derive a general proportionality framework for diversification directly from the procedure presented above, which is described as Algorithm 2.

This framework can be explained as follows. We start with a ranked list S with k empty seats. For each of these seats, we compute the quotient for each aspect t_i following the Sainte-Laguë formula. We then assign this seat to the aspect t_{i^*} with the largest quotient, which marks this seat as a place holder for a document about the aspect t_{i^*}. After that, we need to employ some mechanism to select the actual document with respect to t_{i^*} to fill this seat. Depending on that mechanism, we then need to update the number of seats occupied by each of the aspects t_i accordingly. This process repeats until we get k documents for S or we are out of candidate documents. The order in which each document is put into S determines its ranking. Assuming each document selected for t_i is truly relevant to t_i, the Sainte-Laguë method guarantees proportionality in the final set of documents.

Different choices of document selection mechanisms, which subsequently determine the choices of seat occupation update procedures, will result in different instantiations of our framework. We now present two such instantiations.

Algorithm 2 A Proportionality Framework

```
$s_i \leftarrow 0, \forall i$
for all available seats in the ranked list $S$ do
    for all aspects $t_i \in T$ do
        $quotient[i] = \frac{v_i}{2s_i+1}$
    end for
    $i^* \leftarrow \arg\max_i quotient[i]$
    $d^* \leftarrow$ find the best document with respect to $t_{i^*}$
    $S \leftarrow S \cup \{d^*\}$
    update $s_i, \forall i$ accordingly
end for
```

4.2.2 A Naive Adaptation

We first present a straightforward adaptation from the seat allocation problem above. The Sainte-Laguë method assumes that each member belongs to exactly one party. When a member is assigned to a certain seat, the party naturally takes up the entire seat. Directly applying this technique to our context means assuming each document is associated with a single aspect. As such, we have to determine the aspect for each of the documents $d_j \in R$, which we assume to be the aspect $t_i \in T$ to which d_j is most relevant:

$$\arg\max_{t_i \in T} P(d_j|t_i)$$

As a result, we construct for each aspect t_i a list of documents associated with it in decreasing order of relevance, noted as $M_i = \{d_1^{(i)}, d_2^{(i)}, ..., d_{l_i}^{(i)}\}$ where l_i is the number of documents in M_i. It follows naturally that the best document for an aspect t_i is the first in the list M_i. We refer to this native adaptation as PM-1 and codify it as Algorithm 3.

Algorithm 3 PM-1

```
$s_i \leftarrow 0, \forall i$
for all seats in the ranked list $S$ do
    for all aspects $t_i \in T$ do
        $quotient[i] = \frac{v_i}{2s_i+1}$
    end for
    $i^* \leftarrow \arg\max_i quotient[i]$
    $d^* \leftarrow$ pop $M_{i^*}$
    $S \leftarrow S \cup \{d^*\}$
    $s_{i^*} \leftarrow s_{i^*} + 1$
end for
```

4.2.3 A Probabilistic Interpretation

We now provide a probabilistic interpretation of the Sainte-Laguë method, which removes the naive assumption that a document can only be associated with a single aspect. Instead, we assume all documents $d_j \in D$ are relevant to all aspects $t_i \in T$, each with a probability $P(d_j|t_i)$. This probabilistic interpretation, which we call PM-2, is described by Algorithm 4.

A first point to note is that PM-2 has a different mechanism for document selection. Once a seat is given to the aspect t_{i^*} with the largest quotient, we need to assign to this seat a document that is relevant to t_{i^*}. In the context of multi-aspect documents, however, among several documents all of which are relevant to t_{i^*}, it is sensible to promote documents that may be slightly less relevant to t_{i^*} but are at the same time relevant to other aspects, compared to those that are slightly more relevant to t_{i^*} but are non-relevant to all others. This is, after all, what motivates diversification: we want more users to be able to find what they want. Therefore, PM-2 introduces the parameter λ:

$$d^* \leftarrow \arg\max_{d_j \in R} \lambda \times qt[i^*] \times P(d_j|t_{i^*}) + (1-\lambda) \sum_{i \neq i^*} qt[i] \times P(d_j|t_i)$$

that trades relevance to t_{i^*} with relevance to more aspects. We abbreviate $quotient[i]$ to $qt[i]$ due to the space limitation.

A second point is that when a document d^* is selected for the current seat, since it is assumed to be relevant to all aspects $t_i \in T$, each aspect occupies a certain "portion" of this seat as opposed to a single aspect taking up the entire seat as previously. Intuitively, the degree of occupation of the seat is proportional to the normalized relevance to d^*:

$$s_i \leftarrow s_i + \frac{P(d^*|t_i)}{\sum_{t_j \in T} P(d^*|t_j)}$$

where s_i is the "number", which is now better regarded as "portion", of seats occupied by t_i.

PM-2 can be summarized as a two-step procedure as follows. For each of the k seats in S, it first employs the Sainte-Laguë formula to determine which aspect this seat should go to in order to best maintain the proportionality. Then,

Algorithm 4 PM-2

```
$s_i \leftarrow 0, \forall i$
for all seats in the ranked list $S$ do
    for all aspects $t_i \in T$ do
        $quotient[i] = \frac{v_i}{2s_i+1}$
    end for
    $i^* \leftarrow \arg\max_i quotient[i]$
    $d^* \leftarrow \text{argmax}_{d_j \in R} \;\; \lambda \times quotient[i^*] \times P(d_j|t_{i^*}) + (1-\lambda)\sum_{i \neq i^*} quotient[i] \times P(d_j|t_i)$
    $S \leftarrow S \cup \{d^*\}$
    $R \leftarrow R \setminus \{d^*\}$
    for all aspects $t_i \in T$ do
        $s_i \leftarrow s_i + \frac{P(d^*|t_i)}{\sum_{t_j} P(d^*|t_j)}$
        Since $d^*$ is assumed relevant to all aspects, each of these aspects will take up a certain "portion" of this seat
    end for
end for
```

it selects the document that, in addition to being relevant to this aspect, is relevant to other aspects as well. Finally, it updates the "portion" of seats in S occupied by each of the aspects t_i according to how relevant it is to the selected document.

5. EXPERIMENTAL SETUP

Query and Retrieval Collection. Our query set consists of 98 queries, 50 of which are from the diversity task of the TREC 2009 Web Track (WT-2009) [10] and the other 48 are from TREC 2010 Web Track (WT-2010) [11]. Our evaluation is done on the ClueWeb09 Category B retrieval collection[2], which is also used in both WT-2009 and WT-2010. This collection contains approximately 50 million web pages in English. During query and indexing time, both the query and the collection are stemmed using the Porter stemmer. In addition, we perform stopword removal using the standard INQUERY stopword list.

Baseline Retrieval Model. We use the standard query-likelihood model within the language modeling framework [14] to conduct the initial retrieval run. This run serves both as a mean to provide a set of documents for the diversity models to diversify and a baseline to verify their usefulness. We also use this model as the estimate of relevance $P(d_j|t_i)$ between the document d_j and the aspect t_i.

Spam filtering is known to be an important component of web retrieval [3]. Following Bendersky et al. [3], we use a spam filtering technique as described by Cormack et al. [12] with the publicly available Waterloo Spam Ranking for the ClueWeb09 dataset, which assigns a "spamminess" percentile $S(d)$ to each document d in the collection. In particular, let $p(d_i|q)$ indicate the score the retrieval model assigns to the document d_i, the final score of d_i is given by:

$$P(d_i|q) = \begin{cases} p(d_i|q) & \text{if } S(d_i) \geq 60 \\ -\infty & \text{otherwise} \end{cases}$$

Diversity Models. We evaluate PM-2, the proportionality-aware model we propose for search result diversification. In addition, we will also present results obtained by PM-1 for comparison. Our first diversity baseline model for comparison is MMR [4], which is considered standard in the diversity literature. Since the explicit approach for diversification is generally superior to the implicit approach, we also compare our models to xQuAD, which has been demonstrated to outperform many others in this class [26]. In fact, xQuAD is among the top performers in both diversity tasks of TREC 2009 and TREC 2010 [10, 11]. In addition to these two baselines, we also compare our results to those published by TREC whenever appropriate.

Experiment Design. We use Lemur/Indri [3] to conduct the baseline query-likelihood retrieval run with the toolkit's default parameter configuration. All of the diversification approaches under evaluation are applied on the top-K retrieved documents. All of these models except for PM-1 has a single parameter λ to tune. Readers should refer to the original papers [4, 26] for the interpretation of this parameter in the respective models. We consider for λ values in the range $\{0.05, 0.1, 0.15, ..., 1.0\}$. Our two-fold cross validation enforces complete separation between tuning and testing. In particular, each system is tuned on WT-2009 and tested on WT-2010 and vice versa. We present the result averaged across two folds unless stated otherwise. PM-1, since it is parameter-free, has no tuning involved.

As for the parameter K, we tested $K \in \{50, 100, 500, 1000\}$ and found that all four models achieved their best at $K = 50$. Therefore, all results presented here are achieved with $K = 50$.

Evaluation Metric. We first report our results using CPR, the proportionality metric we propose in Section 3. Since this metric certainly favors our models as they are designed to capture proportionality in the search results, we also report the results of several standard metrics that existing work was designed to optimize. This includes those used in the official evaluation of the diversity task WT-2010 [11]: α-NDCG [7], ERR-IA (a variant of ERR [6]) and NRBP [8]. These measures penalize redundancy at each position in the ranked list based on how much of that information the user has seen and how likely it is that the user is willing to scan down to that position. In addition, we also report our results using Precision-IA [1] and subtopic recall, which indicate respectively the precision across all aspects of the query and how many of those aspects are covered in the results. Last but not least, all of these measures are computed using the top 20 documents each model retrieves also to be consistent with the official TREC evaluation [10, 11].

[2] http://boston.lti.cs.cmu.edu/Data/clueweb09/

[3] http://www.lemurproject.org

Query Aspects. Explicit approaches such as xQuAD and PM-2 assume the availability of the query aspects and their popularity. We first consider the official sub-topics identified by TREC's assessors for each of the queries as its aspects. This simulates the situation where we know exactly what aspects the query has and provides a controlled environment to study the effectiveness of different diversification approaches. As for the aspect popularity, since it is not available in TREC's judgment data, we assume uniform probability for all aspects, which is also consistent with existing work [26, 27, 28].

In order to simulate more practical settings in which we do not know but have to guess the aspects of the query, we follow Santos et al. [26] by adopting suggestions provided by a commercial search engine as aspect representations. However, the search engine is unable to provide suggestions for *four* of the queries in our set. As a result, these experiments are conducted on the subset of 94 queries for which we can obtain aspect representation. We also assume uniform aspect distribution since it was demonstrated to be the most helpful [26].

It is worth noting that the aspects obtained from the search engine certainly do not completely align with the judged aspects provided by TREC assessors. In other words, there will be overlap between the two sets but there will also be generated aspects that are not in the judged set. We will refer to this problem as the misalignment between different sets of aspects and we do not attempt to evaluate the relevance of these misaligned aspects (those that are not in the judged set) in this paper.

6. RESULTS

6.1 Proportionality

In this section, we evaluate how well different methods maintain proportionality in the search results using both TREC sub-topics and suggestions from a commercial search engine as aspect descriptions. Table 1 shows the Cumulative Proportionality score for each system as well as the Win/Loss ratio – the number queries each system improves and hurts respectively. The letters Q, M, X and P indicate statistically significant differences (p-value < 0.05) to Query-likelihood, MMR, xQuAD and PM-1 respectively.

From Table 1, we first notice that although all diversity models are able to provide improvement over the initial retrieval, the magnitude of improvement is very different. The improvement from MMR, for example, is insignificant while the improvement from the other three is more substantial.

Among the four diversity models, PM-2 outperforms all others on both sets of aspects with statistical significance, which demonstrates the effectiveness of our method at capturing proportionality. MMR is the least effective since it is completely unaware of the query aspects, and thus is unable to capture the proportionality among them. xQuAD, on the other hand, does take into account the query aspects to penalize redundancy. For each document selected, xQuAD downweights each of the aspects based on the degree of its relevance to the selected document so that the aspects that have less relevant documents will have higher priority in the next round. Further details about xQuAD can be found in [26]. Hence, xQuAD indeed has the effect of implicitly promoting proportionality, which explains why it significantly outperforms query-likelihood, and also MMR on one of the aspect sets. PM-2, despite the conceptual difference, can be explained using xQuAD's framework of reweighting aspects as well. From this perspective, the biggest difference between the two is that PM-2 uses a more proportionality aware aspect weighting function which is based on the Sainte-Laguë algorithm. This result indeed confirms the effectiveness of this proportionality-driven aspect weighting function.

Table 1: Performance of all techniques in CPR. The letters Q, M, X and P indicate statistically significant differences to Query-likelihood, MMR, xQuAD and PM-1 respectively (p-value $<$ 0.05).

		CPR	Win/Loss
Sub-topics	Query-likelihood	0.4012	
	MMR	0.4018	36/32
	xQuAD	$0.4534_{Q,M}$	48/31
	PM-1	$0.4462_{Q,M}$	46/34
	PM-2	$\mathbf{0.4771}^{X,P}_{Q,M}$	49/33
Suggestions	Query-likelihood	0.3977	
	MMR	0.4016	36/26
	xQuAD	0.4242_{Q}	41/29
	PM-1	0.4067	34/40
	PM-2	$\mathbf{0.4696}^{X,P}_{Q,M}$	48/31

It is worth noting that PM-1, despite being a parameter-free naive version of PM-2, is comparable to xQuAD. There is no statistically significant difference between these two. Comparing PM-1 to PM-2, however, reveals the weakness of its naive assumption. PM-1 associates each of the documents with exactly one aspect, thus it has the risk of associating documents with the wrong aspects. In addition, PM-1 fails to promote documents relevant to multiple aspects. Both of these account for its inferiority to PM-2 with both sets of aspects.

6.2 Standard Redundancy-based Metrics

We now compare our proposed techniques to MMR and xQuAD using standard metrics from the diversity literature as mentioned earlier. Instead of showing the results averaged across two folds, we show the results obtained in each fold separately so that we can compare our results with the official results from TREC. It should be noted that the comparison between our results and those from TREC should be taken as indicative only, since our systems and theirs use different initial retrieval run. Table 2 shows the results for all of the techniques we studied as well as the best performing system on ClueWeb09 Category B reported by TREC. In addition to the scores in each metric, Table 2 also presents the Win/Loss ratio each system achieves over the query likelihood baseline in terms of α-NDCG.

The first observation from Table 2 is that all systems perform worse in all metrics on WT-2009 than they do on WT-2010. The effectiveness of all systems certainly depends on the quality of documents retrieved by the initial retrieval run. Since all of our systems rerank the top 50 returned documents for each query, we examine these documents in both precision-IA and sub-topic recall. The former indicates how many relevant documents for each aspect we have for reranking and the latter indicates how many of the aspects for which we have relevant documents. The results are shown in Table 3, which suggests that the top 50 documents for queries in WT-2009 cover less topics (i.e. many sub-

Table 2: Performance of all techniques in several standard redundancy-based measures. The Win/Loss ratio is with respect to α-NDCG. The letters Q, M, X and P indicate statistically significant differences to Query-likelihood, MMR, xQuAD and PM-1 respectively (p-value $<$ 0.05).

		α-NDCG	Win/Loss	ERR-IA	Prec-IA	S-Recall	NRBP
WT-2009							
Sub-topics	Query-likelihood	0.2979		0.1953	0.1146	0.4327	0.1689
	MMR	0.2963	16/19	0.1922	**0.1221**	0.4447	0.1657
	xQuAD	$0.3300_{Q,M}$	23/15	$0.2207_{Q,M}$	0.1190	**0.4700**	$0.1950_{Q,M}$
	PM-1	0.3076	18/17	0.2027	0.1140	0.4440	0.1738
	PM-2	$\mathbf{0.3473}^{P}$	19/19	$\mathbf{0.2407}^{P}$	0.1197	0.4633	**0.2172**
Suggestions	Query-likelihood	0.2875		0.1895	0.1095	0.4212	0.1634
	MMR	0.2926	16/15	0.1919	0.1108	0.4351	0.1655
	xQuAD	0.2995	14/19	0.1973	0.1089	0.4403	0.1700
	PM-1	0.2870	16/18	0.1830	0.0929^{X}	0.4111	0.1560
	PM-2	**0.3200**	17/19	**0.2139**	$\mathbf{0.1123}^{P}$	**0.4472**	**0.1884**
WT-2009 Best (uogTrDYCcsB) [10]		0.3081	N/A	0.1922	N/A	N/A	0.1617
WT-2010							
Sub-topics	Query-likelihood	0.3236		0.2081	0.1713	0.5479	0.1656
	MMR	0.3349_{Q}	19/14	0.2161	0.1740	0.5694_{Q}	0.1750
	xQuAD	$0.4074_{Q,M}$	29/14	$0.2671_{Q,M}$	0.2028	$0.6410_{Q,M}$	$0.2206_{Q,M}$
	PM-1	$0.4323^{X}_{Q,M}$	32/13	$0.3071^{X}_{Q,M}$	0.1827	$0.6323_{Q,M}$	$0.2654^{X}_{Q,M}$
	PM-2	$\mathbf{0.4546}^{X,P}_{Q,M}$	34/10	$\mathbf{0.3271}^{X}_{Q,M}$	**0.2030**	$\mathbf{0.6503}_{Q,M}$	$\mathbf{0.289}^{X}_{Q,M}$
Suggestions	Query-likelihood	0.3268		0.2131	0.1730	0.5355	0.1722
	MMR	0.3361_{Q}	17/14	0.2206	0.1746	0.5507	0.1819
	xQuAD	$0.3582_{Q,M}$	31/6	$0.2372_{Q,M}$	0.1785	0.5775_{Q}	0.1964_{Q}
	PM-1	0.3664^{X}	25/15	0.2409	0.1654	0.5996	0.1952
	PM-2	$\mathbf{0.4374}^{X,P}_{Q,M}$	33/10	$\mathbf{0.3087}^{X,P}_{Q,M}$	**0.1841**	$\mathbf{0.6279}^{X}_{Q,M}$	$\mathbf{0.2690}^{X,P}_{Q,M}$
WT-2010 Best (uogTrB67xS) [11]		0.4178	N/A	0.2980	N/A	N/A	0.2616

Table 3: Quality of the baseline run for the WT-2009 and WT-2010 query sets in sub-topic recall and precision-IA.

	S-Recall@50	Prec-IA@50
WT-2009	0.54	0.0821
WT-2010	0.7003	0.1486

topics do not get any relevant documents) and also contain considerably less relevant documents for each of the topics than WT-2010. Therefore, there is far less room for improvement on WT-2009 than there is on WT-2010, which leads to the fact that all systems perform better on WT-2010.

Regarding the comparison among diversification techniques, we see a very similar trend as in the previous case with proportionality. In particular, MMR is least effective method due to its lack of awareness of the query aspects. PM-2, on the other hand, outperforms all other methods in almost all metrics with statistically significant improvement in many cases. PM-2 with automatically generated aspects even outperforms the best performing system in TREC evaluation. It should be noted that the best performing system in TREC 2010 of which results we report also use suggestions generated by a search engine as aspect descriptions. This further confirms the effectiveness of PM-2: it provides results with not only a higher degree of proportionality but also a lower degree of redundancy.

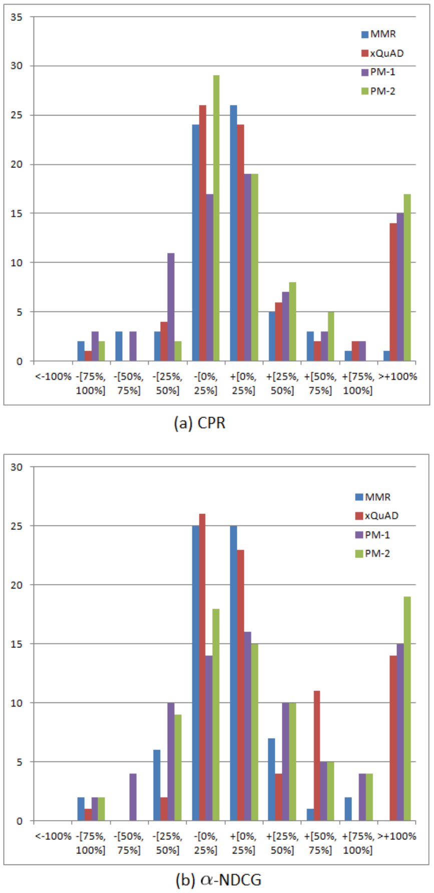
6.3 Improvement Analysis

The analyses in this section are conducted on the entire query set as there is no need to consider WT-2009 and WT-2010 separately. In addition, we only present our analyses with the manually generated set of aspects because we have similar findings with the other set, only to a slightly lesser extent due to the aspect misalignment problem.

We are interested in two aspects of the improvement each technique provides over the initial retrieval: (1) the robustness [21] of the improvement, and (2) the reasons that account for this improvement. Robustness refers to the number of queries each technique improves and hurts together with the magnitude of the performance change. To understand the robustness of each model, we provide a more detailed view of the Win/Loss ratio that was provided earlier in Table 1 and Table 2. In particular, instead of showing how many queries each system improves and hurts over the entire query set, we now look at these numbers with respect to the percentage of the improvement. The histogram in Fig. 1 shows, for various ranges of relative increases (positive ranges) and decreases (negative ranges) in CPR and α-NDCG, the number of queries improved and hurt with respect to the query likelihood baseline.

It can be seen from Fig. 1 that most of the performance changes resulting from using MMR is in the two low ranges $-$[0%,25%] and +[0%, 25%], which indicates that MMR rarely improves or hurts a query drastically. Combined with the fact that it helps and hurts about the same number of queries (35/33), MMR can only provide slight improvement over the baseline.

In contrast, PM-2 and xQuAD provide substantial improvement ($>$ +100%) for several queries. In addition, compared to MMR, these two models hurt about the same number of queries but they improve many more. As a result, PM-2 and xQuAD significantly outperform MMR in most cases. Comparing PM-2 and xQuAD, although they help and hurt about the same number of queries, PM-2 has a much larger magnitude of improvement. This is demon-



(a) CPR

(b) α-NDCG

Figure 1: Robustness of all techniques with respect to the baseline query-likelihood.

strated through Fig. 1 with the fact that PM-2 has more queries in the highest range (>+100%).

The effectiveness of each model depends on two factors: the quality of the initial retrieved set of documents and the model's power to select a diverse subset from that pool of documents. Since all models operate on the same pool, the former factor becomes irrelevant. As for the model power, the key component of both our method and xQuAD is the query likelihood estimate of relevance $P(d_j|t_i)$ between an aspect t_i and a document d_j. While xQuAD uses $P(d_j|t_i)$ to penalize redundancy at every rank, PM-2 uses it to accommodate proportionality. Intuitively, the more accurate $P(d_j|t_i)$ is at telling which document is relevant to which of the aspects of the query, the more diverse the final ranked list will be. In this experiment, we study how well these techniques perform at different level of accuracy $P(d_j|t_i)$ provides.

In order to quantify the accuracy of $P(d_j|t_i)$, we do as follows. Let D be the set of top 50 documents returned for the query q, which has a set of aspects $\{t_1, t_2, ..., t_n\}$. We rank all documents $d_j \in D$ for each of the aspects t_i with $P(d_j|t_i)$ and record the NDCG score. We then use the average of NDCG across all aspects as the measure of accuracy of $P(d_j|t_i)$. Table 4 presents the absolute improvement each model has over the baseline (in both CPR and α-NDCG separately) on different ranges of accuracy of $P(d_j|t_i)$. We also show the number queries and how the baseline query likelihood performs in each of these ranges.

Table 4: CPR and α-NDCG breakdown by ranges of accuracy of $P(d_j|t_i)$.

	$Acc._{P(d_j\|t_i)}$	[0,0.1)	[0.1, 0.2)	[0.2, 0.3)	[0.3, 1.0]
	#q	28	21	26	23
CPR	QL	0.1049	0.4152	0.5296	0.6040
	MMR	−0.0142	+0.0249	−0.0015	−0.0013
	xQuAD	+0.0062	+0.0326	+0.0666	+0.1097
	PM-1	+0.0028	+0.0286	+0.0500	+0.1059
	PM-2	+0.0193	+0.0613	+0.0903	+0.1417
α-NDCG	QL	0.07	0.3078	0.4143	0.4885
	MMR	−0.0079	+0.0146	+0.0133	+0.0013
	xQuAD	+0.0202	+0.0509	+0.0768	+0.0863
	PM-1	+0.0212	+0.0221	+0.0677	+0.1252
	PM-2	+0.0288	+0.0616	+0.1189	+0.1548

Since MMR does not use $P(d_j|t_i)$, its performance obviously does not correlate with the accuracy of $P(d_j|t_i)$. PM-2 consistently provides larger improvement than xQuAD across all ranges of accuracy and metrics. In addition, the gap between the improvement in both CPR and α-NDCG of PM-2 and xQuAD is generally larger as the accuracy of $P(d_j|t_i)$ increases. This clearly indicates using $P(d_j|t_i)$ to optimize proportionality is much more effective than to minimize redundancy, which explains the all-round superiority of PM-2.

In summary, we have demonstrated that MMR is the least effective because it helps and hurts about the same number of queries. PM-2 and xQuAD both help more queries than they hurt, but PM-2 is able to provide substantially larger improvement over the baseline than xQuAD, helping PM-2 to be statistically significantly better than xQuAD even though they both outperform MMR and the baseline. The reason for PM-2's superiority over xQuAD is that PM-2 uses $P(d_j|t_i)$ to accommodate proportionality at every rank, which is more effective than using it to penalize redundancy. The results obtained with PM-2 contain not only a higher degree of proportionality but also a lower degree of redundancy.

6.4 Failure Analysis

Our techniques sequentially go over all "seats" in the result ranked list and decide for each of them which aspect it should go to. After the aspect is determined, PM-1 simply chooses the best document for this aspect according to $P(d_j|t_i)$ while PM-2 might promote documents that are slightly less relevant to this aspect but relevant to other aspects as well.

The problem arises when the initial retrieval fails to find relevant documents for some of the aspects. When a "seat" is assigned to an aspect without relevant documents, $P(d_j|t_i)$ will mistakenly provide some false positive non-relevant documents to fill in that seat, leading to undesirable results. In this section, we will investigate this effect.

Sub-topic recall of the baseline run is certainly the best metric for studying the effect of coverage. Table 5 shows how different systems behave on different ranges of sub-topic recall of the top 50 documents retrieved by the baseline run. For each of the sub-topic recall ranges, Table 5 provides the percentage of queries that each system helps and hurts (marked as "%Q+" and "%Q-" respectively) together with the its relative improvement ("%Imp.") in both CPR and α-NDCG with respect to the baseline. We also show for each range the number of queries as well as the performance of the baseline for references.

We first examine the percentage of queries helped and hurt by each system. At the low recall range ([0,0.5)), both PM-1 and PM-2 hurt more queries than they improve. As the recall goes up, these numbers improve. This trend is especially clear with α-NDCG. This clearly demonstrates the effect sub-topic recall has on our techniques.

xQuAD by its nature does not have the same problem. As a result, the negative effect of low subtopic recall on xQuAD is smaller than it is on our methods: xQuAD has better Win/Loss ratios than both PM-1 and PM-2 on low ([0,0.5)) and medium ([0.5,0.75)) recall range. This helps further explain what we saw earlier in Table 2: although the same techniques perform worse on WT-2009 than they do on WT-2010, PM-1 and PM-2 are the ones with the largest performance difference in terms of Win/Loss ratio. The reason is the subtopic recall of the baseline for WT-2009 is considerably lower than that for WT-2010 (as demonstrated previously in Table 3), which affects our systems the most.

MMR despite not having this problem, it hurts about the same number of queries as our techniques in the low and medium recall ranges due to its overall ineffectiveness. In addition, it helps significantly less queries compared to ours.

With respect to the relative improvement over the baseline, even though xQuAD has better Win/Loss ratios than PM-2 on the low and medium recall ranges, PM-2 still manages to provide larger improvement than xQuAD. Additionally, the gap between the two models becomes substantially larger in the high recall range. This provides additional evidences to support the effectiveness of PM-2.

In summary, even though our proportionality-aware method PM-2 is very effective overall, it depends critically on the coverage of the baseline run.

6.5 Discussion: On Noisy Aspect Descriptions

Given that automatically generated aspects can be helpful for diversification, it is important to know how to generate them. Using query suggestions from commercial search engines in effect is using a "black box" for this important component. Hence, this section aims to provide a preliminary discussion on whether we can use aspects generated by existing work in query suggestion and reformulation for diversification.

While most of those reformulation techniques focus on making user queries more effective [2, 18, 24, 20, 15], some aim to generate reformulations that cover different aspects of the original query [16]. We have adapted these techniques [16] to generate a set of clusters for each of our queries, where each cluster is assumed to represent an aspect of the original query. Details can be found in [16]. We concatenate all queries in each cluster to form a "document", from which we then construct a language model. Finally, we use Indri's weighted query representation of this model as the aspect description and the frequency of the cluster as the popularity of the aspect. The resulting query set consists of 77 queries for which the reformulation technique can provide clusters.

We now re-evaluate all of our techniques using this set of aspects. The results are presented in Table 6. Interestingly, we observe that the performance of xQuAD, PM-1 and PM-2 is substantially lower than with the previous two aspect sets. We observe that this set of aspects, in comparison with the set obtained from the search engine, contains (1) considerably less of the TREC sub-topics and more of other aspects that are not identified by TREC's assesors, and also (2) some unclear aspect descriptions. The low performance of all systems, in fact, clearly demonstrates the aspect misalignment issue and the noisiness of this set.

With these noisy aspects, PM-2 is still better than xQuAD with most of the metrics. The performance of xQuAD is, in fact, even lower than that of the query likelihood baseline in both CPR and α-NDCG. PM-2 still manages to provide improvement, but it is more comparable to MMR. Interestingly, PM-1 is now the best performing approach.

To conclude, aspects generated by the current query reformulation technique are generally not very "effective" for diversification. This notion of "effectiveness", however, has to be taken with care. We have been penalizing all systems for finding documents for aspects that are different to TREC sub-topics. In practice, these unjudged aspects might be relevant to the query as well. This raises the question of how reliable the current evaluation paradigm is that relies on pre-defining a fixed set of aspects for each queries. We will investigate this issue in future work.

7. CONCLUSIONS AND FUTURE WORK

In this paper, we present a different perspective on search result diversification: diversity by proportionality. We consider a result list to be more diverse if the number of documents relevant to each of the aspects is proportional to the overall popularity of that aspect. We then propose Cumulative Proportionality (CPR), an effectiveness measure for proportionality which is based on metrics commonly used for evaluating outcomes of elections. Motivated by the Sainte-Laguë method for assigning seats in a parliament to members of competing political parties, we also present a proportionality-driven framework for diversification. It sequentially determines for each of the "seats" in the result list the aspect that best maintains the overall proportionality with respect to the previously selected topics. It then determines for this seat the best document with respect to that topic. Using this framework, we derive PM-1 – a naive adaptation of the seat allocation mechanism, from which we then develop the probabilistic interpretation, which we called PM-2.

Our results have demonstrated that, with both manually and automatically generated aspect descriptions, PM-2 is statistically significantly better than the top performing redundancy-based technique not only in CPR, but also on several other standard redundancy-based measures. This indicates that promoting proportionality will result in minimal redundancy, as desired by the current standard in diversity.

For future work, we will compare the aspects generated by existing reformulation techniques to the TREC sub-topics in order to quantify the aspect misalignment problem. If many of these misaligned aspects are indeed sensible, we might have to re-examine if predefining a set of topics for each query is a valid strategy for evaluating diversification techniques. In addition, Santos et al. has pointed out that learning to dynamically provide different diversification strategies for different queries based on how ambiguous they are [27] and what intent they have [28] significantly improves the performance of xQuAD. We plan to investigate these approaches since they are potentially beneficial for our model.

Table 5: Performance breakdown by S-Recall of the initially retrieved documents. "%Q+" and "%Q-" indicate respectively the percentage of queries helped and hurt by each technique. "%Imp." indicates the relative improvement of each technique over the baseline query likelihood (QL).

	S-Recall Ranges	[0,0.5)			[0.5,0.75)			[0.75,1.0]		
	#Queries	39			27			32		
		%Q+	%Q-	%Imp.	%Q+	%Q-	%Imp.	%Q+	%Q-	%Imp.
CPR	QL (CPR)	0.2504			0.4689			0.5279		
	MMR	10%	41%	−6%	56%	26%	+4%	53%	28%	+1%
	xQuAD	26%	36%	−6%	74%	19%	+15%	56%	38%	+22%
	PM-1	23%	41%	−5%	63%	30%	+6%	63%	31%	+25%
	PM-2	23%	41%	0%	67%	30%	+18%	69%	28%	+31%
α-NDCG	QL (α-NDCG)	0.1610			0.3791			0.4349		
	MMR	13%	38%	+5%	48%	33%	+3%	53%	28%	+3%
	xQuAD	31%	33%	+6%	74%	22%	+19%	63%	31%	+24%
	PM-1	31%	33%	+5%	56%	37%	+7%	72%	22%	+34%
	PM-2	28%	36%	+8%	67%	30%	+24%	75%	22%	+42%

Table 6: Performance of all techniques with noisy aspect descriptions. Q, M and X indicate significant difference to Query-Likelihood, MMR and xQuAD respectively.

	CPR	α-NDCG	ERR-IA	Prec-IA	S-Recall	NRBP
Query-likelihood	0.3669	0.2637	0.1644	0.1113	0.4107	0.1332
MMR	0.3824	0.2769	0.1722	**0.1353**$_Q$	**0.4450**	0.1387
xQuAD	0.3598	0.2601	0.1620	0.1169$_M$	0.4052	0.1299
PM-1	**0.3943**	**0.2944**$_X$	**0.1961**$_X$	0.1306	0.4189	**0.1685**$_X$
PM-2	0.3703	0.2828	0.1888	0.1157$_M$	0.4010	0.1641

8. ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval, and in part by Vietnam Education Foundation. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

9. REFERENCES

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *Proceedings of WSDM*, pages 5-14, 2009.

[2] R. Baeza-Yates, C. Hurtado and M. Mendoza. Query recommendation using query logs in search engines. In *The ClustWeb Workshop*, pages 588-596, 2004.

[3] M. Bendersky, D. Fisher, and W.B. Croft. UMass at TREC 2010 Web Track: Term dependence, spam filtering and quality bias. In *Proceedings of TREC*, 2010.

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

[5] B. Carterette and P. Chandar. Probabilistic models of ranking novel documents for faceted topic retrieval. In *Proceedings of CIKM*, pages 1287-1296, 2009.

[6] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *Proceedings of CIKM*, pages 621-630, 2009.

[7] C.L.A. Clarke, M. Kolla, G.V. Cormack, O. Vechtomova, A. Ashkan, S. Buttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proceedings of SIGIR*, pages 659-666, 2008.

[8] C.L.A. Clarke, M. Kolla, and O. Vechtomova. An effectiveness measure for ambiguous and underspecified queries. In *Proceedings of ICTIR*, pages 188-199, 2009.

[9] C.L.A. Clarke, N. Craswell, I. Soboroff, and A. Ashkan. A comparative analysis of cascade measures for novelty and diversity. In *Proceedings of WSDM*, pages 75-84, 2011.

[10] C.L.A. Clarke, N. Craswell, and I. Soboroff. Overview of the TREC 2009 Web track. In *TREC*, 2009.

[11] C.L.A. Clarke, N. Craswell, I. Soboroff, and G.V. Cormack. Overview of the TREC 2009 Web track. In *TREC*, 2009.

[12] G.V. Cormack, M.D. Smucker, and C.L.A. Clarke. Efficient and effective spam filtering and re-ranking for large web datasets. Apr 2010.

[13] N. Craswell, O. Zoeter, M.J. Taylor, and B. Ramsey. An experimental comparison of click position-bias models. In *Proceedings of WSDM*, pages 87-94, 2008.

[14] W.B. Croft, D. Metzler, and T. Strohman. Search Engines: Information Retrieval in Practice. *Addison-Wesley*, 2009.

[15] V. Dang and W.B. Croft. Query reformulation using anchor text. In *Proceedings of WSDM*, pages 41-50, 2010.

[16] V. Dang, X. Xue, and W.B. Croft. Inferring query aspects from reformulations using clustering. In *Proceedings of CIKM*, pages 2117-2120, 2011.

[17] M. Gallagher. Proportionality, disproportionality and electoral systems. In *Electoral Studies*, 10(1):33-51, 1991.

[18] R. Jones, B. Rey and O. Madani. Generating query substitutions. In *Proceedings of WWW*, pages 387-396, 2006.

[19] A. Lijphart. Electoral systems and party systems: A study of twenty-seven democracies, 1945-1990. *Oxford University Press*, 1994.

[20] Q. Mei, D. Zhou and K. Church. Query suggestion using hitting time. In *Proceedings of CIKM*, pages 469-477, 2008.

[21] D. Metzler and W.B. Croft. Latent concept expansion using markov random fields. In *Proceedings of SIGIR*, pages 311-318, 2007.

[22] D. Rafiei, K. Bharat and A. Shukia. Diversifying web search results. In *Proceedings of WWW*, page 781-790, 2010.

[23] F. Radlinski and S. Dumais. Improving personalized web search using result diversification. In *Proceedings of SIGIR*, pages 691-692, 2006.

[24] X. Wang and C. Zhai. Mining term association patterns from search logs for effective query reformulation. In *Proceedings of CIKM*, pages 479-488, 2008.

[25] J. Wang and J. Zhu. Portfolio theory of information retrieval. In *Proceedings of SIGIR*, pages 115-122, 2009.

[26] R. L. T. Santos, C. Macdonald, and I. Ounis. Exploiting query reformulations for web search result diversification. In *Proceedings of WWW*, pages 881-890, 2010.

[27] R. L. T. Santos, C. Macdonald, and I. Ounis. Selectively diversifying web search results. In *Proceedings of CIKM*, pages 1179-1188, 2010.

[28] R. L. T. Santos, C. Macdonald, and I. Ounis. Intent-aware search result diversification. In *Proceedings of SIGIR*, pages 595-604, 2011.

[29] C. Zhai, W.W. Cohen, and J. Lafferty. Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. In *Proceedings of SIGIR*, pages 10-17, 2003.