Spring 2014

A Proportionality-based Approach to Search Result Diversification

Van Bac Dang
University of Massachusetts - Amherst, vdang@cs.umass.edu

A PROPORTIONALITY-BASED APPROACH TO SEARCH RESULT DIVERSIFICATION

A Dissertation Presented

by

VAN B. DANG

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2014

School of Computer Science

A PROPORTIONALITY-BASED APPROACH TO SEARCH RESULT DIVERSIFICATION

A Dissertation Presented

by

VAN B. DANG

Approved as to style and content by:

W. Bruce Croft, Chair

James Allan, Member

Benjamin M. Marlin, Member

Michael L. Lavine, Member

Lori A. Clarke, Chair
School of Computer Science

To my family

ACKNOWLEDGMENTS

This thesis would not have been possible without the great support from my advisor, W. Bruce Croft. He taught me an immense number of lessons in conducting research. He monitored my progress closely and provided me with specific instructions when I was inexperienced. He gave me a great deal of flexibility in choosing what I wanted to do as I became more mature. He always had great advice for me when I needed it. He was highly involved in the process of writing papers. Not only did he help revise my drafts, he also made sure that his edits are clear. His revisions helped improve my writing skill greatly. I know that Bruce dedicated a huge amount of time and effort to training me and I am extremely thankful for everything that he has done.

I would also like to thank James Allan. Although I did not work with James directly, he had given me many good pieces of advice about how to do well in graduate school. James is also one of my committee members. His comments and encouragement are extremely helpful. I would also like to thank my other committee members: Benjamin M. Marlin and Michael L. Lavine. Their criticisms have made this work better in many ways.

I am very thankful for the incredible support that I received from all the staffs of CIIR and the School of Computer Science at UMass. Specifically, I would like to thank Kate Morruzzi for always having the solution for whatever problems I have, Dan Parker for his excellent technical support, Glenn Stowell for making sure that I got my paychecks regularly, Jean Joyce for doing the important work behind the scene, and Leeanne Leclerc for helping me so much with my enrollment and especially my graduation.

I must also thank all my CIIR friends and colleagues: Laura Dietz, Mostafa Keikha, David Fisher, Michael Bendersky, Elif Aktolga, Xiaobing Xue, Marc Cartright, Henry Feild, Jinyoung Kim, Jeff Dalton, Sam Huston, Tiger Wu, Jae-Hyun Park, Youngho Kim, Chia-Jung Lee, Weize Kong, I. Zeki Yalniz, Shiri Dori-Hacohen, Xing Yi, Niranjan Balasubramanian, Tamsin Maxwell, Ethem Can and everyone else. You always gave me valuable feedback for my practice talks, which made the real ones much better. I learned a lot from all of you. I know that I will miss all the conversations we had. You make my time at CIIR unforgettable.

While at CIIR, I had the opportunity to spend a summer with Giridhar Kumaran and Adam Troy at Bing. This internship gave me a better understanding of how it was like to do research in the industry. I graciously thank Giridhar and Adam for the great experiences that I had.

Before joining CIIR, I was fortunate to have Le Dinh Duy, Duong Anh Duc, Ho Bao Quoc and Dong Thi Bich Thuy as my advisors at the University of Science, Vietnam. I thank them for teaching me the very first things about academic research, believing in me, and supporting my decision to pursue graduate study abroad. I also want to thank Akiko Aizawa from the National Institute of Informatics, Japan, for giving me great research advice and experiences while I was an intern there.

Last but not least, I would like to thank my family. I am forever grateful to my parents, Dang Ky Sach and Diep Cam Huy, for always giving me the love and the guidance that I need, and for their continuous encouragement for me to pursue graduate study in the United States. I must also thank my grand parents, my aunt and uncles – Diep Song Nguyen, Duong Dieu Xoa, Diep Van Huy, Diep Hy Luong and Diep Hy Tai – for their financial support when I was in high school and college. Finally, I would like to thank my wife, the one and only "mini Van" (Van Nguyen), for believing in me. This work would not have been possible without her understanding, caring and patience. I am sincerely indebted to "mini" for her love and generosity.

This work was funded in part by a grant from the Vietnam Education Foundation (VEF). The opinions, findings, and conclusions stated herein are those of the author and do not necessarily reflect those of VEF.

This work was supported in part by the Center for Intelligent Information Retrieval, in part by IBM subcontract #4913003298 under DARPA prime contract #HR001-12-C-0015, in part under subcontract #19-000208 from SRI International, prime contractor to DARPA contract #HR001-12-C-0016, in part by NSF CLUE IIS-0844226, in part by NSF grant #IIS-0534383, and in part by NSF grant #IIS-0711348. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect those of the sponsor.

ABSTRACT

A PROPORTIONALITY-BASED APPROACH TO SEARCH RESULT DIVERSIFICATION

MAY 2014

VAN B. DANG

B.Sc., UNIVERSITY OF SCIENCE, HO CHI MINH CITY, VIETNAM

M.Sc., UNIVERSITY OF MASSACHUSETTS AMHERST

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor W. Bruce Croft

Search result diversification addresses the problem of queries with unclear information needs. The aim of using diversification techniques is to find a ranking of documents that covers multiple possible interpretations, aspects, or topics for a given query. By explicitly providing diversity in search results, this approach can increase the likelihood that users will find documents relevant to their specific intent, thereby improving effectiveness.

This dissertation introduces a new perspective on diversity: diversity by proportionality. We consider a result list more diverse, with respect to some set of topics related to the query, when the ratio between the number of relevant documents it provides for each of these topics matches more closely with the topic popularity distribution. Consequently, we derive an effectiveness measure based on proportionality and propose a new framework for optimizing proportionality in search results, which we show to be more effective than existing techniques.

Diversification would be impractical without the ability to automatically infer the set of topics associated with the user queries. Therefore, we study cluster-based techniques for generating these topics from publicly available data sources.

Based on the challenges that we observe with topic generation, we present a simplified term-based representation for query topics. Specifically, we propose to identify for each query a single set of terms that describes its topics. This set is provided to a diversification technique which in effect treats each of the terms as a topic to determine coverage in the search results. We call this approach term level diversification and we show that it can promote diversity with respect to the topics underlying the input terms. This simplifies the task of finding a set of query topics, which has proven difficult, to finding only a set of terms. We also present a technique as well as several data sources for generating these terms effectively.

TABLE OF CONTENTS

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

CHAPTER 1

INTRODUCTION

User queries do not always clearly represent the actual information needs. Existing research estimated that 16% of the queries submitted to search engines are ambiguous with multiple possible interpretations (R. Song et al., 2007; Clough et al., 2009). For example, people searching for "TREC" might be looking for the Text Retrieval Conference, Texas Real Estate Commission or Tropical Research and Education Center. Even when different users share a common interpretation, their information needs might still be different (Clarke et al., 2008). While one user may be interested in publications at the TREC conference, others might want to know about different tracks being organized at this conference. It is unclear which of these two aspects a particular user is interested in.

Traditional relevance-based retrieval models (Robertson & Walker, 1994; Ponte & Croft, 1998; Liu, 2009) do not take into account such diversity in user information needs. Instead, they assume that there is a *single implicit* topic associated with all relevant documents for each query. Thus, their objective is to retrieve documents that are potentially relevant to this topic. Document relevance can be modeled using various signals ranging from simple textual matches between documents and queries (Robertson & Walker, 1994; Ponte & Croft, 1998) to more complicated user behavior (Agichtein et al., 2006). Regardless, these models have a high risk of retrieving a ranking with no or insufficient coverage for the topics that the user desires due to its single-topic assumption.

Search result diversification was introduced to address this short-coming. These techniques attempt to return a ranking of documents that covers multiple topics of the query. Topics denote the multiple possible information needs, intents, interpretations or aspects 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.

Intuitively, a diverse result list is one that can address the diverse information needs underlying the user queries effectively. How to quantify such notion, however, is unclear. Therefore, researchers make reasonable interpretations about what makes a diverse ranking effective. They mostly rely on the notion of *novelty* and *redundancy*. These are derived from the understanding of user behaviours in a web search environment, known as the cascade model (Craswell et al., 2008). Users are assumed to examine the result list top down and eventually stop reading because either they have found what they need or they have run out of patience. 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. A ranked list is considered more diverse if it contains more novelty, or equivalently, less redundancy. Enforcing maximal novelty at every position in the ranking ensures coverage for more topics. It also induces a document ordering where documents from more topics are surfaced to higher ranks, reducing the user effort involved in finding them. The relative popularity of the query topics must also be considered. The main idea is that the best document for the more popular topics should be presented before those for the less popular topics, and that there should be more documents for topics that are more popular.

Consequently, the task of finding a diverse ranking of documents, with respect to the topics of the query, has been studied primarily from this perspective of maximizing *novelty*. Maximizing coverage alone has been shown to be NP-hard (Agrawal et al.,

2009). As a result, existing models (Carbonell & Goldstein, 1998; Zhai et al., 2003; Agrawal et al., 2009; Carterette & Chandar, 2009; Santos et al., 2010a) are all greedy approximations. They construct the ranking by iteratively selecting documents from some pool of candidate documents. At each iteration, they pick the document with maximal novelty with respect to those that have been selected for earlier ranks. Similarly, evaluation measures (Clarke et al., 2008; Clarke, Kolla, & Vechtomova, 2009; Chapelle et al., 2009) also focus on penalizing result lists with high redundancy. We will discuss these techniques and measures as well as their differences in more detail in Chapter 2.

Although the novelty approach considers the popularity of query topics, it does not explicitly maintain a ratio of documents to present for each topic. In this dissertation, we propose a new perspective on diversity which enforces this ratio. We call this *diversity by proportionality*. It means that we want the ratio of documents returned for each topic to match the relative topic popularity as closely as possible. Consider an example scenario where we have to return ten documents for the query "TREC", which we assume to have two topics, *the TREC conference* and *Texas Real Estate*, whose popularity is 80% and 20% respectively. We hypothesize that a result list with eight documents about the conference and two documents about real estate is more representative of the overall interest in these topics, thus more effective, than a list with five documents for each topics. Proportionality naturally implies coverage. Enforcing maximal proportionality at every position in the result list helps surface documents for more topics, which has the same effect as promoting novelty. It also guarantees that the more popular topic will be represented before less popular topics. Under this perspective, we propose a measure for proportionality as well as a model that can provide a proportional ranking of documents with respect to a set of query topics and their popularity. Our experimental results have shown that document

rankings provided by our technique not only have higher proportionality but also lower redundancy compared to existing methods.

We have been discussing diversification with respect to a set of query topics without mentioning where these topics come from. The success of diversification techniques has been observed mostly with topics that are readily available. For example, the related searches or suggested queries that commercial search engines provide for a query can be used to represent its topics. While their effectiveness for diversification has been confirmed (Santos et al., 2010a), how to generate these topics is unclear. Alternatively, the Open Directory Project (ODP) taxonomy, a human-edited hierarchy of categories that aims to describe the web has also been used as a source of topics (Agrawal et al., 2009). Each query is classified to one or several categories in this taxonomy using a random walk algorithm (Fuxman et al., 2008). These category descriptions are then used for diversification. While ODP provides topics of high quality, because they are created by human editors, it potentially has a coverage issue. Enriching such a resource could be expensive.

Generating a set of topics for a query automatically, on the other hand, has had rather limited success. Although there have been several attempts to do this, their effectiveness for diversifying web search results has yet been fully evaluated. For example, Carterette and Chandar (2009) apply Latent Dirichlet Allocation (Blei et al., 2003) on a set documents retrieved for a query to recover its topics. Alternatively, they also cluster these documents with k-nearest neighbor and use a language model estimated from each cluster to represent a topic. Although their methods have proven useful, their evaluation is done only on a small newswire collection. Although there are more recent studies focusing on web corpora (Dou et al., 2011; Zheng et al., 2011) which achieve promising results, their success is often tied to certain parameter values chosen manually for the model. If these values are identified automatically, via cross validation for instance, it is unclear how that affects the performance.

In this dissertation, we propose a technique for automatically inferring the topics for any user query from publicly available resources. In particular, we consider each related query, or reformulation, a possible representation of a query topic. We study different clustering techniques for grouping topically similar reformulations together to form a detailed description for each of the query aspects. Our results indicate that many of the generated reformulations correspond to the true query topics identified by human judges. Furthermore, our clusters are topically consistent and effective for diversification.

Since human usually describes a topic with a coherent group of terms, existing techniques for topic generation model topics as groups of terms (Dou et al., 2011; Zheng et al., 2011) or distributions of terms (Carterette & Chandar, 2009; He et al., 2012). Because this approach has not been entirely fruitful, we propose a less strict topic representation in which we directly model terms without their topical grouping. What this means is that instead of generating a set of two topics for the query *TREC – trec conference* and *texas real estate commission* – we only attempt to find a limited set of terms such as: *trec*, *conference*, *texas*, *real*, *estate* and *commission*. Each term will be treated as a topic and the term set is provided as input to existing techniques for diversification as though they were a conventional topic set.

We demonstrate, using a ground-truth set of query topics, that diversification with our term-based representation, which we call *term level diversification*, can be as effective as its topic-based counter part. The reason that this works lies in the nature of the diversification methods, which we will elaborate later on. We then show that these terms can be generated automatically using a relatively simple existing method proposed by Lawrie and Croft (2003) for document summarization. These automatically generated terms are substantially more effective for diversification than the full topic structures generated by existing methods.

The key point of our term level approach is that the term set must contain terms that can describe multiple topics of the query. As such, finding these terms is a diversification problem by itself. Therefore, we apply existing document diversification techniques, as well as our proposed proportionality method, to this term diversification problem and investigate their effectiveness.

1.1 Contributions

The contributions of this thesis are as follows.

1. We introduce 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 relative popularity among these topics. This is in contrast with the existing view, which focuses on promoting novelty in the result ranking. Consequently, we propose an effectiveness measure for proportionality called Cumulative Proportionality.

2. We propose a framework for optimizing proportionality for search result diversification. We demonstrate that our method is more effective than the top performing approach in the current literature not only according to our proportionality measure, but also using several standard redundancy-based metrics.

3. We propose a cluster-based method to automatically infer query topics from publicly available resources: anchor text and the Microsoft N-gram service. We compare the effectiveness of different clustering techniques as well as similarity measures.

4. We introduce a term-based representation of query aspects for diversification. Instead of modeling a set of aspects, each of which is a group or a distribution of terms, we directly model terms without their topical grouping. This effectively

reduces the key problem of recovering a set of topics to finding a set of terms, which is potentially a simpler problem.

5. We show that this term set can be generated automatically and effectively using an existing technique for document summarization (Lawrie & Croft, 2003). Since finding these terms is also a diversification problem, we also apply several techniques that have been proposed for diversifying documents to diversify terms. Furthermore, we explore the use of multiple resources including retrieved documents, wikipedia, freebase, anchor text and query logs in this term generation process.

1.2 Dissertation Outline

In Chapter 2, we survey the related work on search result diversification. This includes retrieval techniques, topic generation methods, and evaluation measures. We will also discuss the standard TREC dataset, which is the basis of our empirical evaluation.

In Chapter 3, we first introduce our proportionality perspective on diversification. We then derive our measure for it, which we call Cumulative Proportionality (CPR). In addition, we present our framework for maximizing proportionality in search results.

In Chapter 4, we study the correlation between CPR and existing diversity measures which are based on the notion of redundancy. We then present the results of retrieval experiments with our proportionality framework for diversification. We compare it with existing approaches using both our proportionality measure and a variety of standard redundancy-based metrics.

In Chapter 5, we describe how anchor text and the Microsoft N-gram service, both of which are publicly available, can be used to infer query topics for diversification.

We then describe our technique, which leverages these two sources, in more detail and then evaluate it.

In Chapter 6, we first present the intuition behind our term level diversification approach and explain how it works. We then present the document summarization technique (Lawrie & Croft, 2003) that we use to automatically generate these terms. After that, we explain how term generation can be considered a diversification problem itself and the connection between this summarization technique and document diversification methods. Finally, we present the sources of information that can be used for term generation.

In Chapter 7, we evaluate our term level approach to search result diversification using both human-created terms that describe a set of ground-truth query topics as well as terms that are generated automatically. We compare our approach to the conventional topic level method. We also study the effectiveness for term generation of the document diversification techniques as well as the resources we explorer in the previous chapter.

Finally, in Chapter 8, we summarize the findings of this dissertation and address some promising future directions.

CHAPTER 2

BACKGROUND AND RELATED WORK

One important theoretical statement about retrieval effectiveness is known as the *Probabilistic Ranking Principle* (PRP) (Robertson, 1997). It states that a ranking is most effective when its documents are presented in order of decreasing probability of *relevance* to the user who submitted the query. This implies the *document independence assumption*: the relevance of a document is independent of other documents in the ranking.

The PRP does not specify the notion of relevance. Instead, it is up to specific retrieval models to interpret and estimate it. The common interpretation shared by many models is that there is a single implicit topic underlying all relevant documents for each user query. The objective is thus to score all documents in the collection independently using some estimate of their relevance to this topic and rank them by this score. This interpretation leads to several standard models such as BM25 (Robertson & Walker, 1994) and Query Likelihood (Ponte & Croft, 1998), which estimate relevance based on how well the document text matches the query terms. Additional evidence beyond textual matching (i.e. user behaviours (Agichtein et al., 2006)) can also be incorporated to improve the reliability of this estimate via the learning to rank framework (Liu, 2009).

Regardless of how relevance is estimated, these methods have two limitations. Firstly, the single-topic assumption is unrealistic. In practice, user information needs are seldom clear based on the initial query that they specify. Existing research estimated that 16% of web queries are ambiguous with multiple possible interpretations

(R. Song et al., 2007; Clough et al., 2009). This means that, search results should allow for coverage of multiple topics.

Secondly, the document independence assumption is not realistic. If a user finds a document in the ranking relevant, all lower-ranked documents providing the same information will become less valuable to this user. This has been recognized for a long time (Goffman, 1964) and recently been confirmed by the cascade model of user behavior (Craswell et al., 2008). This model suggests that users typically examine the result list from top to bottom and the probability that they click on each document decreases as they click on documents at earlier ranks.

Due to these two assumptions, it is entirely possible that traditional retrieval models generate a result list with too many documents on one topic while leaving others uncovered. This results in a high risk of having no or insufficient coverage for the topics in which the users are interested. To reduce this risk, one has to promote topical diversity in the search results. This task is known as *search result diversification*.

We start this chapter by explaining the current notion of topical diversity in search results, which we will refer to as *diversity* for brevity. We then survey existing techniques for search result diversification. After that, we provide a brief description of related work for automatically inferring query topics. Next, we explain the standard diversity evaluation process including datasets and effectiveness measures. Finally, we briefly discuss related work on non-topical search result diversification as well as other related areas.

2.1 Diversity and Search Result Diversification

Diversity is a generic concept that appears in many areas. For example, diversity among university students often indicates the range of countries or cultures from

which the students come. In ecology, diversity reflects the distribution of plants or animals across species.

Diversity in search results refers to how well these results address the diverse information needs, topics, or aspects underlying the user query. Intuitively, this might involve several factors. Firstly, how many topics or aspects should be covered? Secondly, how many documents should be retrieved for each topic? Additionally, how should these documents be ordered? Some of these questions are easier to answer than others. For example, while we certainly want to have coverage for more query topics, there might not be a "right" answer to how many documents from different topics should be interleaved or how they should be ordered.

As a result, one needs to make reasonable interpretations about what makes a diverse ranking effective. One interpretation is based on the notion of *coverage* (Agrawal et al., 2009; Carterette & Chandar, 2009). This view ignores the order of documents in the search results. It assumes that if a search engine returns 10 documents to the users, they will examine all of them. As a result, an effective diverse ranking is one that has at least one document for each of the query topics.

Another interpretation of diversity, which take into account document ordering, is based on *novelty* and *redundancy* (Clarke et al., 2008; Clarke, Kolla, & Vechtomova, 2009; Chapelle et al., 2009). This is derived from the cascade model of user behavior (Craswell et al., 2008). Since users are believed to examine the result lists from the top down, a document at any rank providing the same information as those at earlier ranks can be considered *redundant*. Likewise, a *novel* document is one that provides information that has not been covered by any of the previous documents. The novelty of a ranking is accumulated from the novelty of its documents in a rank-dependent fashion. A result list with maximal novelty (or equivalently, minimal redundancy), which loosely implies maximal novelty at every rank, is considered more effectively diverse. It is important to point out that *novelty* subsumes *coverage*. If

we put together a ranking from all documents in the collection, the result ranking with maximal novelty naturally has maximal coverage. Briefly put, a result list with maximal novelty is a ranking with maximal coverage (with respect to the query topics) in which the documents are ordered in such a way that minimizes the average rank where all users will find a result that is useful to them.

As a result, diversification has been studied from the perspective of maximizing novelty, or equivalently, minimizing redundancy (Carbonell & Goldstein, 1998; Zhai et al., 2003; Santos et al., 2010a). As we shall see later, even the techniques that emphasize coverage (Agrawal et al., 2009; Carterette & Chandar, 2009) achieve their goal by promoting novelty.

Searching through the entire document collection for a ranking of documents that maximizes some utility, however, is impractically expensive. It is also unnecessary since many documents will not be relevant to the user information needs. As a result, diversification is performed using re-ranking. First, a relevance-based retrieval model (e.g. Query Likelihood (Ponte & Croft, 1998)) is used to obtain an initial ranking of documents that are potentially relevant to some of the topics underlying the user query. This ranking is then re-ordered to maximize either coverage or novelty with respect to some set of topics associated with this query.

Formally, let $T = \{t_1, t_2, ..., t_n\}$ indicate the set of topics underlying the query q. Let $R = \{d_1, d_2, ..., d_m\}$ be the initial ranking of documents retrieved for q. The diversification task is to select and rank k documents ($k \leq m$) from R to form a diverse ranked list S that maximizes novelty with respect to T.

2.2 Diversification Techniques

One of the effectiveness criteria of a diverse ranking is topic coverage. Unfortunately, even if we ignore document ordering, finding a document set with maximal coverage is NP-hard. This can be proved by showing that this task is equivalent to

the maximum coverage problem, a known NP-hard problem (Agrawal et al., 2009). Since the greedy approach has been proven to achieve the best approximation factor for the maximum coverage problem (Nemhauser et al., 1978; Feige, 1998), all diversification techniques – coverage-based and novelty-based alike, are polynomial-time greedy algorithms as outlined in Algorithm 1. This greedy framework iteratively selects documents in R to put into S. At each iteration, it chooses the document d^* such that $S \cup \{d^*\}$ has maximum utility $f(q, d, S)$, which can be coverage or novelty.

Interestingly, although the coverage view of diversity does not take into account the order among documents, the techniques derived from it conveniently provide this ordering due to their greedy nature. One can see from Algorithm 1 that what they do can be explained as promoting coverage of new topics (with respect to the documents selected earlier) at every rank. This makes the coverage-based approach (Agrawal et al., 2009; Carterette & Chandar, 2009) virtually identical to the novelty-based one (Carbonell & Goldstein, 1998; Zhai et al., 2003; Santos et al., 2010a) in terms of optimization objective since promoting coverage for new topics is equivalent to promoting novelty.

Algorithm 1 The greedy approach to search result diversification

```
1: procedure DIVERSIFY(q, R)
2:     S ← ∅
3:     while |S| < Min(k, |R|) do
4:         d* ← argmax_{d∈R} f(q, d, S)
5:         S ← S ∪ {d*}
6:         R ← R \ {d*}
7:     end while
8:     return S
9: end procedure
```

Although the utility function function $f(q, d, S)$ can be referred to more specifically as the novelty function, we will retain the term utility function for the following reason. In an ideal world where all documents in R are relevant, this utility can be solely the novelty of each document with respect to those that have been selected

earlier. In practice, however, this ranking usually contains several non-relevant documents as well. Favoring novel documents alone likely ends up promoting non-relevant documents since they usually provide "novel" information compared to the relevant ones. As a result, this utility function $f(q, d, S)$ usually captures both the relevance of each candidate document to the query and its novelty. How to estimate novelty as well as how to combine it with relevance is where existing methods differ.

How existing methods estimate novelty is determined primarily by whether or not they explicitly represent the query topics in their model. As a result, they are often categorized as being *implicit* or *explicit*.

2.2.1 The Implicit Approach

As the name implies, this approach does not explicitly represent query topics within their models. Instead, it assumes each document has its own latent topics, reflected by its vocabulary. Novelty is promoted by selecting the document at each iteration that uses the most different vocabulary compared to those selected earlier, as given by some similarity measures.

The pioneer technique in this area is known as Maximal Marginal Relevance (MMR) (Carbonell & Goldstein, 1998). This technique was originally proposed to reduce redundancy in document rankings as well as in text summarization and has become the canonical baseline for diversification since then. Using the greedy framework, it scores each candidate document by its relevance to the user query discounted by its maximum similarity with respect to the documents that have been selected earlier:

$$f_{MMR}(q, d, S) = \lambda R(d, q) - (1 - \lambda) \max_{d_j \in S} Sim(d, d_j)$$

where $R(d, q)$ indicates how relevant d is to q and $Sim(d, d_j)$ is the document similarity function. While cosine was used in the original paper, any other measures should be applicable.

Motivated by the idea of MMR, Zhai et al. (2003) propose to model novelty, which they described as *dependent relevance*, using the language modeling approach:

$$f_{LM}(q, d, S) = R(d, q)(1 - \lambda - P(d|S))$$

where $R(d, q)$ could be any probabilistic relevance estimate (such as query likelihood (Ponte & Croft, 1998)) and $P(d|S)$ is an estimate of the probability that the words in d come from the language model of the documents in S. Lower $P(d|S)$ indicates more novelty.

Inspired by the portfolio theory in finance (Markowitz, 1952), Wang and Zhu (2009) proposed a coverage-based approach to maximize the expected relevance of a ranking, i.e. the average estimated relevance score of its documents, at a specified variance [1]. Documents with similar vocabulary usually have similar relevance estimate. Therefore, having a certain degree of variance in the ranking promotes documents with different vocabulary, thereby increasing novelty. This objective is achieved by the same greedy approach that is very similar to MMR except that document similarity is modeled by Pearson's correlation.

Chen and Karger (2006) used a similar utility function as that of Zhai et al. (2003) except for the novelty estimate. Instead of comparing each candidate document d with those in S, this approach models d's novelty as the probability that d is relevant to the user information need conditioned on the fact that those in S are assumed to be non-relevant. They show that this maximizes the probability that there is at least one document in the result ranking that is useful to the user. Note that this method achieves diversity as an "unplanned" effect and has not been formally evaluated using standard redundancy measures.

[1] A similar approach is introduced independently by Rafiei et al. (2010)

2.2.2 The Explicit Approach

Intuitively, the relative popularity of query topics should have some effect on how search engines order their result documents. The implicit approach, however, has very little control over which topics of the query to present in the results, which is decided entirely by the document similarity measure being used. Therefore, it is unclear how the relative topic popularity can be incorporated into such models.

The explicit approach was proposed to address these short comings. It explicitly maintains a set of query aspects and their importance or popularity, and attempts to return documents for each of these aspects, ordered in a way that maximizes novelty (Carterette & Chandar, 2009; Agrawal et al., 2009; Santos et al., 2010a). The difference between existing techniques in this class is also in the way they model novelty and combine it with relevance.

It is important to note that these techniques assume that the set of topics $T = \{t_1, t_2, ..., t_n\}$ is available for the user query q. We treat the task of generating these topics as a separate problem which we will discuss later in Section 2.4.

Agrawal et al. (2009) used the Open Directory Project (ODP) taxonomy to model query topics. Each query q is mapped to a small set of topics T in this taxonomy using a random walk algorithm (Fuxman et al., 2008), which also provides a distribution specifying the probability $P(t|q)$ that q belongs to each of the topics $t \in T$. Their diversification technique, IA-Select, defines $f(q, d, S)$ as follows:

$$f(q, d, S)_{IA-Select} = \sum_{t \in T} P(d|t) \times P(t|q) \times \prod_{d_j \in S} (1 - P(d_j|t))$$

where $P(d|t)$ is some probabilistic estimate of the relevance of d to the topic t. $f(q, d, S)$ denotes the marginal utility of d, which is interpreted as the probability that that d satisfies the user where all of the documents that come before it fail to do so. The objective of this greedy approach can be explained in different ways.

From the user modeling perspective, if a user examines the document at position i, it means all $i-1$ documents IA-Select provided so far are on the wrong topics with respect to this user. As a result, the document to put in this i-th position should be the one that has the most potential for satisfying this user conditioned on the fact that all previous $i-1$ documents did not.

Operationally, it could be explained as follows. Let $\bar{P}(S|t) = \prod_{d_j \in S}(1 - P(d_j|t))$, which indicates the probability that all documents in S fail to satisfy the topic t_i. Note that $\bar{P}(S|t_i)$ decreases as any document d is put into S since every document is relevant to t with some probability. The exact amount it is decreased depends on the relevance of d. At any iteration, more documents in S that are more relevant to t means lower $\bar{P}(S|t)$. Therefore, $\bar{P}(S|t)$ can be regarded as an indicator of how well the topic t is currently covered, or satisfied, by S – higher values means less well covered. As a result, by selecting the document d^* with maximum $f(q, d, S)$ at every step of the greedy framework, IA-Select favors those documents that can satisfy the topics that have not been well satisfied. This is how novelty is promoted.

Following a different line of reasoning, Santos et al. (2010a) arrived at the same objective function. They further interpolate it with the relevance of the candidate document with the query:

$$f(q, d, S)_{xQuAD} = (1 - \lambda)P(d|q)) + \lambda \sum_{t \in T} P(d|t) \times P(t|q) \times \prod_{d_j \in S} (1 - P(d_j|t))$$

The main difference between xQuAD and IA-Select is their assumption about query topics. While Agrawal et al. (2009) uses ODP to represent query topics , Santos et al. (2010a) consider the related queries obtained from commercial search engines for each query as its topics. This essentially determines how $P(d|t)$ is estimated. As far as diversification techniques are concerned, however, they are very similar.

Carterette and Chandar (2009) proposed a probabilistic set-based technique that is similar to IA-Select in the objective of maximizing topic coverage in the result

ranking. Their derivation, however, leads a different utility function:

$$f(q, d, S)_{Set} = \prod_{t \in T} \left(1 - (1 - P(d|t)) \prod_{d_j \in S} (1 - P(d_j|t)) \right)$$

This utility aims to capture the probability that $S \cup \{d\}$ contains at least one document for each of the query topic. Maximizing this utility ensures that their method always selects the document for the topics that have not been (well) covered by those currently in S. In other words, it promotes novelty at every position in the result ranking.

More recently, Zheng et al. (2012) proposed a family of coverage-based utility functions that linearly combines the relevance of the document to the query and with its novelty in a similar fashion as xQuAD. They show that some of these can improve the novelty in the result ranking.

In summary, all of the above mentioned methods are very similar in that they all employ a greedy framework that attempts to select the most novel document at every step. In contrast, we introduce a new framework for diversification that is based on maximizing proportionality. Though we only provide two instantiations of this framework, further derivation is certainly possible. Our proportionality model also assumes a set of query topics, which makes it an explicit approach.

It is worth noting that, while most of the above techniques are unsupervised, there is some effort in applying machine learning to diversification as well. Yue and Joachims (2008) propose to use structural SVM to predict how well a candidate set of documents satisfies the query topics based on how well it covers the vocabulary in the input ranking, which is assumed to have sufficient coverage for those topics. The effectiveness of this approach, however, is unclear since it is only evaluated in the ideal setting where the input ranking to be diversified does not contain non-relevant documents. In addition, it requires expensive editorial judgment data (i.e. identifying

query topics and judging which documents are relevant to which of these topics). To avoid editorial judgment, online learning has also been considered where the models are learned and updated on-the-fly as implicit feedback is collected from users (Raman et al., 2011, 2012; Radlinski et al., 2008; Slivkins et al., 2010; Yue & Guestrin, 2011).

2.3 Automatic Generation of Query Topics

Compared to the implicit approach for diversification, the explicit approach offers more control over which topics to cover in the search results. The effectiveness of the explicit approach, however, has been observed primarily with topics that are either created manually (e.g. ODP (Agrawal et al., 2009)) or obtained directly from related queries provided by commercial search engines (Santos et al., 2010a). Although there have been quite a few techniques for generating query topics automatically, not all of them are designed with diversification effectiveness in mind. Therefore, the success of the diversification techniques with topics that are generated automatically is rather limited.

Carterette and Chandar (2009) propose to cluster a set of documents retrieved for a query using k-nearest neighbor. They then construct a relevance model (Lavrenko & Croft, 2001) from each cluster and use it to represent a query topic. Alternatively, they also apply Latent Dirichlet Allocation (Blei et al., 2003) on the same set of documents to obtain these topics. Their method has proven effective on a small newswire collection. It is unclear if these results can be generalized to noisy web collections.

Radlinski et al. (2010) seek topics from a search engine log. In particular, they propose to cluster related queries in a large proprietary log and use each cluster to represent a query topic. They show that this approach can provide topically consistent clusters of queries. As a result, it was adopted for the topic development procedure at TREC (Clarke, Craswell, & Soboroff, 2009; Clarke et al., 2010; Clarke, Craswell,

Soboroff, & Voorhees, 2011; Clarke & Craswell, 2012). Human judges examined the clusters for each user query and decided the topics in which searchers are interested. They then manually created the description for each of these topics. Nevertheless, the effectiveness of these generated clusters for diversification has yet to be confirmed.

Inspired by the work of Radlinski et al. (2010), we propose to cluster related queries generated from anchor text and web ngrams. Both of these resources are publicly available whereas query logs are not. We show that diversification with topics generated using our method outperforms a standard relevance-based retrieval model. Furthermore, He et al. (2012) show that combining topics generated using our method and those generated from the retrieved documents and query logs further improves diversification effectiveness.

Ma et al. (2010) present a technique that generates a diverse set of suggestions for a given query by performing a diversified random walk on a click graph starting from the input query. Alternatively, Y. Song et al. (2011) apply the learning to rank approach to generate these suggestions that leverages not only click data but also session information and retrieval results. Since these techniques are designed for query suggestion, their effectiveness for diversification has not been studied.

Instead of relying on a single source of information, Dou et al. (2011) propose to combine multiple sources. In particular, they extract and score anchor text and queries from a search log that contain all of the query terms. The top ranked anchors and queries are used as topics. Furthermore, they cluster documents initially retrieved for the query based on key phrases in their snippets and use these clusters as additional topics. Although their evaluation shows that diversifying search results using these topics can improve diversity, the number of topics extracted from each of the sources to be used is selected manually. It is thus unclear if these topics remain effective in a fully automatic setting.

To the best of our knowledge, only more recent methods show consistent improvement over non-diversification retrieval baselines on web corpora. This includes the technique by He et al. (2012) and the one by Santos et al. (2013). While the former infers query topics from multiple sources of information using the regularized topic modeling approach (Cai et al., 2008), the latter uses learning to rank to select related queries from a query log. The difference between this work and that by Y. Song et al. (2011) is in the features they use.

It can be seen that existing techniques represent a topic by a distribution of terms (Carterette & Chandar, 2009), a query (Ma et al., 2010; Y. Song et al., 2011; Santos et al., 2013), some anchor text (Dou et al., 2011), or a cluster of queries (Radlinski et al., 2010). All of these can be regarded as a coherent group of terms. We argue that while this representation has its advantage, such as being human readable, it might not be necessary for diversification. In this dissertation, we experiment with a simpler representation for a set of query topics. Instead of modeling this set with each topic being a group of terms, we model these terms directly without their grouping structure. We show that being able to identify the these topic terms, without their topical grouping, is sufficient to improve diversity in search results. This effectively reduces the task of finding a set of topics into finding an appropriate set of terms. We then show how these terms can be generated effectively.

2.4 Diversity Evaluation

2.4.1 TREC Corpus

The Text REtrieval Conference (TREC) has created numerous reusable test collections over the years for several retrieval tasks. These test corpora have been extensively used by information retrieval researchers to evaluate their models and to ensure the reproducibility of their experimental results published in academic conferences.

The first test collection for diversity evaluation is developed for the aspect retrieval task in the TREC 6-8 Interactive tracks (Over, 1997, 1998; Hersh & Over, 1999). This corpus includes 20 queries, each of which is associated with a number of aspects. Since these aspects are identified by human assessors, this corpus does not reflect the genuine needs of real users who issue the query. This issue is then addressed in the diversity task in the TREC Web tracks 2009-2012 (Clarke, Craswell, & Soboroff, 2009; Clarke et al., 2010; Clarke, Craswell, Soboroff, & Voorhees, 2011; Clarke & Craswell, 2012), in which both the queries and their aspects are identified from a search log, which reflects genuine user information needs.

In this dissertation, we use the TREC corpus developed for the diversity task in the 2009-2012 Web tracks. The retrieval collection is ClueWeb09 category B, which contains approximately 50 million web pages with highest crawled priority derived from a large general English-language web corpus. This corpus is composed of 200 topics which represents the information needs that users might have given this web collection. Each of these topics consists of (1) a short query intended to be used for retrieval experiments, (2) a longer description describing this topic in more detail and (3) a set of sub-topics corresponding to different interpretations, facets, or aspects the query might have. On average, there are 4.12 sub-topics per query with a variance of 1.3. The smallest and largest number of sub-topics a query has is 2 and 8 respectively.

These sub-topics are identified from a query log in a semi-automatic fashion. For each query, they find other queries that are likely to co-occur in the same session. The same expansion step is run again to obtain an even larger set of related queries, which might contain both on-topic and off-topic ones. After that, a graph is constructed from these related queries where two nodes are connected if they have clicks on the same URL. This helps filters out the off-topic queries. Finally, agglomerative clustering is run on the graph. The result clusters were presented to human assessors who then determined the set of sub-topics for this query and provided a description

for each sub-topic. In this dissertation, we refer to these "sub-topics" as "topics" of the query.

Each query is categorized as either "ambiguous" or "faceted". "Ambiguous" queries are those with multiple possible interpretations (their topics thus correspond to different interpretations) and "faceted" queries are those whose intent is under-specified (their topics denote different aspects). In general, all existing diversification techniques can work without distinguishing the two types of queries. Similarly, we do not differentiate between them. We treat both interpretations and facets of a query as its topics.

To assess how effective a system is at providing a diverse result ranking for the 200 queries, this test collection also provides a set of documents that are manually judged as either relevant or non-relevant to each of the query topics. The set of query topics as well as the relevance judgments are used as the ground-truth for the purpose of automatic evaluation. In the next section, we describe how evaluation is conducted in more detail.

2.4.2 Diversity Measures

The novelty-based approach to diversity is clearly demonstrated through several standard effectiveness metrics such as ERR-IA (Chapelle et al., 2009), α-NDCG (Clarke et al., 2008), and NRBP (Clarke, Kolla, & Vechtomova, 2009). These measures have the same general form as follows:

$$Div.@k = \sum_{r=1}^{k} \frac{1}{Discount(r)} \sum_{t \in T} p_t \times Relevance(d_r, t) \times Redundancy(d_r | d_1, ..., d_{r-1})$$

where d_r is the document at rank r in the ranking being measured, T is the set of topics and p_t is the importance of a topic $t \in T$. The general idea is that they give a reward for every document in the list that is relevant to at least one of the topics, weighted by the importance of the corresponding topics. Since a good ranking is

one in which relevant documents are presented at earlier positions, the reward for these documents are discounted by a function of their rank. To promote novelty, they further discount this reward based on the redundancy of this document. Therefore, a ranking with more novel (and relevant) documents at earlier ranks will receive a higher effectiveness score.

The difference between these measures is in they way they compute relevance, redundancy and the rank-dependent discount, which are determined by their interpretations of what make a ranking effective.

ERR-IA (Expected Reciprocal Rank - Intent Aware) (Chapelle et al., 2009) measures the expected (reciprocal) rank at which the user will find useful information. It is computed as:

$$ERR\text{-}IA@k = \sum_{r=1}^{k} \frac{1}{r} \sum_{t \in T} p_t \times R_r^t \times \prod_{j=1}^{r-1} (1 - R_j^t)$$

where R_j^t is the relevance of the j-th document to the topic t, which is calculated as:

$$R_j^t = \frac{2^{R_j} - 1}{2^{R_{max}}}$$

where R_{max} is the highest scale of the graded relevance judgments. The redundancy of a document that is relevant to some subset of topics $T' \subseteq T$ is measured by the the relevance of the documents before it to T'.

α-NDCG ($\alpha-$ Normalized Discounted Cumulative Gain) (Clarke et al., 2008) extends the traditional NDCG measure (Järvelin & Kekäläinen, 2002), which accumulates the reward for each relevant document discounted by its rank, by further discounting this reward by the redundancy of the document being considered:

$$\alpha\text{-}NDCG@k = \sum_{r=1}^{k} \frac{1}{log_2(r+1)} \sum_{t \in T} J_r^t \times (1 - \alpha)^{\sum_{j=1}^{r-1} J_j^t}$$

where J_j^t is the binary relevance of the j-th document to the topic t and $\alpha \in (0, 1]$ indicates the probability that the assessor is making a mistake by judging a document

as relevant (to any of the topics). α is often set to 0.5 in TREC official evaluations. Similar to ERR-IA, α-NDCG measures redundancy of a document by the relevance of the documents that come before it in the ranking. Unlike ERR-IA, α-NDCG ignores relative topic importance, which is its weakness. This is not an issue for evaluation on the TREC corpus, however, since topics are assumed to be equally important.

NRBP (Novelty- and Rank-Bias Precision) (Clarke, Kolla, & Vechtomova, 2009) extends the RBP metric (Moffat & Zobel, 2008) to measure the expected number of relevant documents a user encounters when scanning the result list:

$$NRBP@k = N \times \sum_{r=1}^{k} \beta^{k-1} \sum_{i \in I} \frac{p_i}{|A_i|} \sum_{a \in A} J_r^a \times (1-\alpha)^{\sum_{j=1}^{r-1} J_j^a}$$

where I is the set of interpretations the query has and A_i is the set of aspects under the interpretation i. They attempt to address the general case where a query can have multiple interpretations, each of which has several aspects. Their assumption is that each user is only interested in one interpretation while many of the aspects under this interpretation might be of interest. The rest is similar to α-NDCG. The additional parameter β^k indicates the probability that the user continues reading after examining the k-th document. It is said to model user patience and is assumed constant across k.

The three measures above are often referred to as the cascade measures since they are derived from the cascade user model (Craswell et al., 2008). There are other measures that were not designed with respect to this user model. Examples include the family of *intent aware* measures (Agrawal et al., 2009). Each of them is a linear combination of a traditional measure of relevance Λ (e.g. Λ could be NDCG (Järvelin & Kekäläinen, 2002), MAP, etc.) computed independently for each of the query topics:

$$\Lambda\text{-}IA@k = \sum_{t \in T} p_t \Lambda_t$$

They are designed to address the case when queries are strictly ambiguous and a user is never interested in more than one topic. Thus, it aims to capture the expected relevance of the results across users. Another example is subtopic recall S-Recall (Zhai et al., 2003), which is the fraction of query topics to which a result list contains at least one relevant document. S-Recall, in fact, is equivalent to the instance recall measure used in the aspect retrieval task in the TREC 6-8 Interactive tracks (Over, 1997, 1998; Hersh & Over, 1999). It is important to point out that although $ERR\text{-}IA@k$ fits mathematically into this family, it is a cascade measure. Alternatively to the intent-aware family, Sakai and Song (2011) present a different way to extend the traditional relevance measure that they call the $D\#$-measure family.

Although there has been some effort to compare these measures (Clarke, Craswell, Soboroff, & Ashkan, 2011; Ashkan & Clarke, 2011), it is found that many of them have high correlation with user preferences and there is no significant difference in their predictive power (Sanderson et al., 2010). As a result, the common practice is to conduct evaluation using multiple measures, as has been done by the official TREC evaluation for several years (Clarke, Craswell, & Soboroff, 2009; Clarke et al., 2010; Clarke, Craswell, Soboroff, & Voorhees, 2011; Clarke & Craswell, 2012). Another reason for this practice is that different measures capture different features of the results. For example, while subtopic recall (Zhai et al., 2003) indicates the topic coverage of a result list, intent aware precision (Agrawal et al., 2009) provides more insight into how many relevant documents there are for each topic. ERR-IA (Chapelle et al., 2009) reflects the ranking of documents within each topic as well as an all-round diversity score.

In this dissertation, we propose a new measure which aims to capture the proportionality in search results. It is intended to provide insights on a different aspect of effectiveness. As such, we suggest using it with existing redundancy-based measures.

Last but not least, the measures above are designed to capture diversity in a ranked list, which is specific to the context of information retrieval. Therefore, they are different from diversity measures in other areas, which are mainly set-based. For instance, the Simpson index (Simpson, 1949) in ecology, which is the probability that two individuals taken at random come from the same species, has no notion of ranking.

2.5 Non-Topical Search Result Diversification

Instead of providing a document ranking with coverage for multiple query topics, recent work has proposed to perform diversification with respect to other dimensions of the queries that are non-topical. One of such dimensions is sentiment (Demartini, 2011; Kacimi & Gamper, 2011; Aktolga & Allan, 2013; Aktolga, 2014). Consider a query with a controversial topic: "abortion". It is likely that there will be people who support it (positive), people who are against it (negative), and those who with a neutral attitude towards it (neutral). Therefore, in response to such queries, search engine should return results with coverage for all three sentiments. This will provide the searchers with a better overview of the topic.

Another dimension for diversification is temporal (Keikha et al., 2012; Berberich & Bedathur, 2004; Aktolga, 2014). Consider a query with an ambiguous temporal profile (Jones & Diaz, 2007): "earthquakes in Turkey". Since there have been several earthquakes, providing search results with coverage for these events that happened at different points of time is arguably more helpful to the users' understanding on this subject.

The primary focus of the work in this area is to point out that there are other important dimensions (other than topical) that search result diversification should take into account. As far as techniques are concerned, on the other hand, prior work in sentiment and temporal diversification mostly employ methods that are proposed

for topical diversification, which we have presented earlier in this chapter. In other words, they also employ a greedy framework to select documents that are different to those that have been selected in the previous iterations. The measures of the difference between documents are based on their time (which can be the time when the documents were published or when the events they describe occurred) or sentiments, as opposed to the topics they describe.

2.6 Other Related Research Areas

Novelty and redundancy, which are fundamental to the current notion of diversity, have been studied in other tasks. One of such tasks is the *new event detection* problem, which is a part of the Topic Detection and Tracking (TDT) initiative (Allan, Carbonell, et al., 1998). The task is to find new events (or topics) in a stream of events. An event is considered new if it is different to those before it in the stream. Although this task is related to diversification in that a sub-stream that consists of only new events is highly diverse, its focus is on finding these new events and not on the diversity of the results. Therefore, some of the techniques (Allan, Papka, & Lavrenko, 1998; Allan et al., 2001) are similar to those for diversification, but the goals are ultimately different. In addition, many techniques for new event detection (Allan, Papka, & Lavrenko, 1998; Allan et al., 2001) rely heavily on temporal clues that are not guaranteed to be present in web documents. Furthermore, they do not consider the popularity of the events while many diversification techniques take into account the relative topic popularity (Agrawal et al., 2009; Santos et al., 2010a).

Another task that is also concerned with novelty is *novel sentence detection*, which has been investigated at the TREC 2002-2004 novelty tracks (Harman, 2002; Soboroff & Harman, 2003; Soboroff, 2004). Given a ranking of sentence-segmented documents for a user query, a system's task is to first identify and filter out the non-relevant *sentences*. It then revisits the ranking consisting of the remaining sentences and

discards those that do not contain any new information with respect to those at the earlier ranks. Although this task is related to diversification, the notion of novelty at the sentence level is at a much finer granularity than document-level novelty. This brings different challenges (i.e., isolating relevant sentences is very difficult; and a system's ability to find novel sentences depends critically on how good it is at identifying relevant sentences (Allan et al., 2003)) and thus different solutions to the forefront.

It is worth noting that although this problem of finding a sentence ranking with maximal novelty is formulated more or less as a filtering problem, there are approaches that attempt to re-order the sentences in the input document list to provide a new ranking with a higher degree of novelty (Larkey et al., 2002; Allan et al., 2003). These techniques, in fact, were inspired by MMR (Carbonell & Goldstein, 1998). Specifically, they iteratively select sentences that are different to those selected in the previous iterations using various measures of differences. Since MMR has been applied for diversification, these methods can be considered diversification techniques as well.

CHAPTER 3

DIVERSITY BY PROPORTIONALITY: AN ELECTION-BASED APPROACH

3.1 Introduction

As mentioned in the first two chapters, the problem of finding a diverse ranked list of documents has been investigated mainly from the perspective of promoting *novelty* or penalizing *redundancy*. Existing techniques promote diversity by penalizing result lists with too many documents on the same topic, reducing the redundancy of coverage (Agrawal et al., 2009; Carterette & Chandar, 2009; Santos et al., 2010a). Similarly, many standard effectiveness measures for diversity (Clarke et al., 2008; Clarke, Kolla, & Vechtomova, 2009; Chapelle et al., 2009) are also based on this notion of novelty.

We approach the same task from a different perspective. We view the problem of finding a good result list of any given size, with respect to the topics or aspects of the query, as the task of finding a representative sample that reflects the overall user interest in these topics. Using a simple (and well-worn) example, for a query "java", 90% of the time people click on the web pages about programming language and 10% on documents about the island. We hypothesize that a result list containing ten documents where only one of them was about the island would be more representative, thus more effective, than a result list containing five documents on each topic.

Although it appears that promoting proportionality results in an abundant number of documents for the programming topic, we argue that this is necessary for two reasons. First, consider the alternative where a system does not respect proportionality. In the case that a query has more than ten topics, it is possible that the topic

that most users care about is not covered in the top ten documents returned by this system. Second, the users who are interested in the programming topic might be looking for information on different sub-topics such as Java tutorials, advanced Java programming and the differences between Java and C/C++. By providing sufficient representation for this topic of programming, we enable the possibility to further diversify the results with respect to its sub-topics. Although we have not done this in this dissertation, we believe that it will increase the likelihood of satisfying a larger portion of the 90% of users whose interest is in Java programming.

This notion of *proportionality* of ours is independent of how user interest is measured. As an alternative to using click distribution, one can also think of the number of documents on a topic in a retrieved set or collection as a reflection of user interest. Consequently, we treat the problem of finding a diverse ranking of documents as finding a *proportional representation* with respect to a distribution of user interest.

Finding a proportional representation is a critical part of most electoral processes. The problem is to assign a set of seats in a 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 topic of the query as a "party" and the topic 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.

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

It sequentially visits each "seat" in the list and determines for each of them to which topic it should be allocated in order to maintain proportionality. Then it selects the best document for the selected topic to occupy this "seat". In addition, we also present a new effectiveness measure that captures proportionality in search results.

In the following sections, we will first formally describe our notion of proportionality and how to measure it (Section 3.2). We then present our proportionality-based framework for diversification (Section 3.3). In the next chapter, 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 novelty-based metrics including α-*NDCG* (Clarke et al., 2008), *ERR-IA* (Clarke et al., 2008) and *NRBP* (Clarke, Kolla, & Vechtomova, 2009) that existing work has been designed to optimize. This indicates that optimizing search results for proportionality leads an diverse result list with low redundancy.

3.2 Proportionality

In this section, we will first explain the notion of *proportionality* in the context of information retrieval. We then describe our effectiveness measure for it.

3.2.1 Definition of Proportionality

Let $T = \{t_1, t_2, ..., t_n\}$ indicate a set of topics for a query q and $P = \{p_t | t \in T\}$ indicate the distribution of user interest over these topics. Additionally, let S be some *set* of documents that are relevant to at least one of the query topics, and $P' = \{r_t | t \in T\}$ denote a distribution where r_t is the fraction of documents in S that are relevant to the topic t. We define the *proportionality* of S with respect to T as the similarity between its P' and P. S is considered *perfectly proportional* if and only if P' is identical to P. As P' diverges from P, under some measure of divergence, S

is considered less proportional. We will present a measure for divergence in the next section.

Let us revisit the example in the previous section, in which the popularity of the topic about "java" programming language is 90% and the popularity of the topic about an island named "java" is 10%. Let $\{x, y\}$ denote any set of documents with x documents about programming and y documents about the island. In this case, $\{9, 1\}$ is perfectly proportional. Using a standard least square distance, while $\{8, 2\}$ is not perfectly proportional, it is more proportional than $\{7, 3\}$.

Let S now represent some *ranking* of documents and S_k represent the *set* of top k documents of S. We define the proportionality of S as the average proportionality of all S_k with $k \in [1, |S|]$. The idea is that a ranking is more proportional if the set of documents it provides at every rank is more proportional. Enforcing high proportionality at every rank helps surface relevant documents from different topics at early positions in the result list, thereby reducing the user effort involved in finding them, while making sure topics are represented sufficiently and in a reasonable order.

3.2.2 Effectiveness Measure

The notion of proportionality is 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 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 (P) each party receives and the percentage of seats it gets (P'). Among those, the least square index (LSq)

(Gallagher, 1991) is one of the standard metrics for measuring dis-proportionality (divergence of P' from P):

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

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 an example of the proportionality of a retrieved *set* of *ten* documents for the query *satellite*, which we assume to have two topics: *satellite radio* and *satellite phone* with equal popularity of 50%. Due to the possible presence of non-relevant documents, we have to create a third "topic" to account for non-relevant documents. As a result, proportionality requires this list to contain *five* relevant documents for each of the two topics and *zero* documents for the non-relevant "topic". 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 topics of a query. It then is possible that a topic can be "rewarded" with additional documents while others still have as many relevant documents as they deserve. For example, let us assume we have found nine documents for the query *satellite* in which five of them are about *phones* and four of them are about *radio*. Although the tenth

document should be about *radio*, it does not hurt to get one that is relevant to both topics even though our results will then overly represent the *phone* topic slightly.

Second, it is undesirable for any party to get any more seats than it deserves. This is not true in our case due to the presence of the non-relevant "topic". Overly representing an actually query topic is not as bad as overly representing the non-relevant topic. Using the same example above, although selecting a document about *satellite phones* as the tenth search result is not ideal, it is better than choosing 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 topics. *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, dis-proportionality at rank K, calculated as follows:

$$DP@K = \sum_{topic} \sum_{t \in T} c_t (v_t - s_t)^2 + \frac{1}{2} n_{NR}^2 \tag{3.1}$$

where v_t is the number of relevant documents that the topic t should have, s_t is the number of relevant documents the system actually found for this topic, n_{NR} is the number of non-relevant documents this system retrieved, and

$$c_t = \begin{cases} 1 & v_t \geq s_t \\ 0 & \text{otherwise} \end{cases}$$

Formula (3.1) has two important properties. The first is that it penalizes a result set for under-representing any topic of the query ($s_t < v_t$) but not for over-representing them ($s_t > v_t$), which addresses the first issue associated with *LSq*. The second is that while the over-representation of a query topic is not penalized, the over-representation of the non-relevant "topic" ($n_{NR} > 0$) 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, *Max-DP* is given by:

$$Max\text{-}DP@K = \sum_{topic} \sum_{t \in T} v_t^2 + \frac{1}{2}K^2$$

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

$$PR@K = 1 - \frac{DP@K}{Max\text{-}DP@K}$$

Following our definition of proportionality for a ranking, the *Cumulative Proportionality (*CPR*)* measure for rankings is calculated as follows:

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

Table 3.1 shows how *CPR* is calculated for two rankings R_1 and R_2 returned for the query *satellite* with two topics – *satellite radio* and *satellite phone* – of equal popularity. To compute *CPR@1* for the ranking R_1, notice that the right "number" of documents for both topics is $v_{radio} = v_{phone} = 0.5$. Since the first document in this ranking is about *satellite phone*, $s_{phone} = 1$ and $s_{radio} = 0$. It follows that *PR@1=0.75*. To compute *PR@2*, note that $v_{radio} = v_{phone} = 1$. Furthermore, $s_{phone} = s_{radio} = 1$ and $n_{NR} = 0$ since there is one relevant for each topic and there are no non-relevant documents. This set of two documents is perfectly proportional, thus *PR@2=1*. One can proceed in a similar fashion and obtain $CPR@5 = 0.94$. Repeating the same computation for the ranking R_2 yields $CPR@5 = 0.81$. Although the two rankings contain the same number of relevant documents for each topic, R_1 has higher *CPR@5* since it is more proportional at every rank.

Table 3.1: *CPR* computation for two ranking R_1 and R_2. v_{radio} and v_{phone} indicate the minimum number of documents the ranking must have for the *phone* and *radio* topic respectively. s_{radio} and s_{phone}, on the other hand, indicate the number of documents the ranking actually has on these two topics. n_{NR} is the number of documents that are not relevant to either topic

	Rank	Topic	v_{radio}	s_{radio}	v_{phone}	s_{phone}	n_{NR}	*DP*	*M.-DP*	*PR*
R_1	1	*phone*	0.5	0	0.5	1	0	0.25	1	0.75
	2	*radio*	1	1	1	1	0	0	4	1
	3	*radio*	1.5	2	1.5	1	0	0.25	9	0.97
	4	*phone*	2	2	2	2	0	0	16	1
	5	non-relevant	2.5	2	2.5	2	1	1	25	0.96
CPR@5 $= 0.94$										
R_2	1	*phone*	0.5	0	0.5	1	0	0.25	1	0.75
	2	*phone*	1	0	1	2	0	1	4	0.75
	3	non-relevant	1.5	0	1.5	2	1	2.75	9	0.69
	4	*radio*	2	1	2	2	1	1.5	16	0.91
	5	*radio*	2.5	2	2.5	2	1	1	25	0.96
CPR@5 $= 0.81$										

3.3 A Proportionality Framework for Diversification

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 3.2.2. We then demonstrate the analogy between this and our problem of proportionality-based diversification, which helps us derive our technique from this method.

3.3.1 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 2. 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. Note that we assume that $|M_i|$ is larger than the total number of seats available.

Algorithm 2 The Sainte-Laguë method for seat allocation

```
1: s_i ← 0, ∀i
2: for all available seats in the parliament do
3:     for all parties P_i do
4:         quotient[i] = v_i / (2s_i + 1)
5:     end for
6:     k ← arg max_i quotient[i]
7:     m* ← the best member of P_k
8:     Assign the current seat to m*
9:     M_k ← M_k \ {m*}
10:    s_k ← s_k + 1
11: end for
```

3.3.2 Diversity by Proportionality

3.3.2.1 Framework

Let q indicate the a query, $T = \{t_1, t_2, ..., t_n\}$ indicate the topics for q and p_t indicate the popularity of a topic $t \in T$. In addition, let $R = \{d_1, d_2, ..., d_m\}$ be the ranked list of potentially relevant documents returned for q by some standard retrieval models and $P(d|t)$ indicate some estimate of the probability that the document $d \in R$ is relevant to the topic $t \in T$. The task is to select a subset of R to form a diverse ranked list S of size k.

As mentioned earlier, existing techniques (Agrawal et al., 2009; Carterette & Chandar, 2009; Santos et al., 2010a) generally favor an S with smaller redundancy. Our idea, on the other hand, is to favor a ranking S with higher proportionality. This objective is, in fact, very similar to that of the seat allocation problem above. As a result, by substituting the notion of "party" for "topic" and "votes" (v) for

"popularity" (p), we derive a general proportionality framework for diversification directly from the procedure presented above, which is described in Algorithm 3.

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 q_t for each topic t following the Sainte-Laguë formula. We then assign this seat to the topic t^* with the largest quotient, which marks this seat as a place holder for a document about the topic t^*. After that, we need to employ some mechanism to select the actual document with respect to t^* to fill this seat. Depending on that mechanism, we then need to update the number of seats occupied by each of the topics t 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 is truly relevant to t, for each rank $r \in [1, k]$, the Sainte-Laguë method will select the document that maximizes the proportionality of the set of top r documents in S, which makes S a highly proportional ranking by definition.

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 3 A Proportionality Framework

$s_i \leftarrow 0, \forall i$
for all available slots in the ranked list S **do**
    for all topics $t \in T$ **do**
        $q_t = \frac{p_t}{2s_t+1}$
    end for
    $t^* \leftarrow \arg\max_{t \in T} q_t$
    $d^* \leftarrow$ *find the best document with respect to* t^* *from* $R \setminus S$
    $S \leftarrow S \cup \{d^*\}$
    update $s_t, \forall t$ *accordingly*
end for

3.3.2.2 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 entire seat is taken up. Directly applying this technique to our context means assuming each document is associated with a single topic. Therefore, we have to determine the topic for each of the documents $d \in R$, which we assume to be the topic $t \in T$ to which d is most relevant:

$$\arg\max_{t \in T} P(d|t)$$

As a result, we construct for each topic t a list of documents associated with it in decreasing order of relevance, noted as $M_t = \{d_1^{(t)}, d_2^{(t)}, ..., d_{|M_t|}^{(t)}\}$. It follows naturally that the best document for a topic t is the first in the list M_t. We refer to this native adaptation as `PM-1` and codify it in Algorithm 4.

Algorithm 4 PM-1

$s_t \leftarrow 0, \forall t \in T$
for all slots in the ranked list S **do**
    for all topics $t \in T$ **do**
        $q_t = \frac{p_t}{2s_t+1}$
    end for
    $t^* \leftarrow \arg\max_{t \in T} q_t$
    $d^* \leftarrow$ pop M_{t^*}
    $S \leftarrow S \cup \{d^*\}$
    $s_{t^*} \leftarrow s_{t^*} + 1$
end for

3.3.2.3 A Realistic Interpretation

We now provide a more realistic interpretation of the Sainte-Laguë method, which removes the naive assumption that a document can only be associated with a single topic. Instead, we assume all documents $d \in D$ are relevant to all topics $t \in T$, each with a probability $P(d|t)$. This interpretation, which we call `PM-2`, is described by Algorithm 5.

Algorithm 5 PM-2

$s_t \leftarrow 0, \forall t \in T$
for all slots in the ranked list S **do**
    for all topics $t \in T$ **do**
        $q_t = \frac{p_t}{2s_t+1}$
    end for
    $t^* \leftarrow \arg\max_{t \in T} q_t$
    $d^* \leftarrow \text{argmax}_{d \in R} \ \lambda \times q_{t^*} \times P(d|t^*) + (1-\lambda) \sum_{t \neq t^*} q_t \times P(d_j|t)$
    $S \leftarrow S \cup \{d^*\}$
    $R \leftarrow R \setminus \{d^*\}$
    for all topics $t \in T$ **do**
        $s_t \leftarrow s_t + \frac{P(d^*|t)}{\sum_{t' \in T} P(d^*|t')}$
        Since d^* is assumed relevant to all topics, each of these topics will take up a certain "portion" of this seat
    end for
end for

A first point to note is that `PM-2` has a different mechanism for document selection. Once a seat is given to the topic t^* with the largest quotient, we need to assign to this seat a document that is relevant to t^*. In the context of multi-topic documents, however, among several documents all of which are relevant to t^*, it is sensible to promote documents that are also relevant to other topics over those that are only relevant to t^*. This is, after all, a general goal of diversification: we want to have broader topic coverage so that more users to be able to find what they want. Our proportionality approach has an additional goal: we want to return more documents for the more popular topics. One way to achieve these goals is to score each document based on its relevance to all topics, weighted by the quotient of each topic:

$$d^* \leftarrow \arg\max_{d \in R} \ q_{t^*} \times P(d|t^*) + \sum_{t \neq t^*} q_t \times P(d|t) \tag{3.2}$$

The problem with this formulation is that it does not put sufficient emphasis on covering t^*, which is required to maximize proportionality. It is possible that a document that is relevant to other topics but not t^* is selected over the ones that are relevant to

t^*. For example, consider a query q which has six topics with equal popularity. Let us assume that the first document selected d_1 is relevant only to the first four topics:

- d_1: $P(d_1|t_i) = 0.3, \forall i \in [1..4]$, $P(d_1|t_5) = 0$ and $P(d_1|t_6) = 0$ [2]

One can verify that the quotient for the four three topics is 0.11 and the quotient for the fourth topic is 0.17. Thus, the next document to select should cover t_4 and t_5 to maximize proportionality.

Now let us assume there are three candidate documents to choose for the second position in S:

- d_2: $P(d_2|t_6) = 0.3$, and $P(d_2|t_i) = 0, \forall i \in [1..5]$
- d_3: $P(d_3|t_6) = 0.3$, $P(d_3|t_5) = 0.3$, and $P(d_3|t_i) = 0, \forall i \in [1..4]$
- d_4: $P(d_4|t_i) = 0.3, \forall i \in [1..4]$, $P(d_1|t_5) = 0$ and $P(d_1|t_6) = 0$

Although it is hard to argue generally which of these candidates is the best choice for users, d_2 and d_3 are clearly better for making the ranking S more proportional since they t_5 and/or t_6. While d_4 is relevant to many topics as well, choosing d_4 does not improve proportionality at all. Between d_2 and d_3, d_3 is better since it provides larger improvement in proportionality. Ideally, we need a scoring function that can assign the highest score to d_3.

Applying Formula (3.2), however, the score for d_2 is 0.026, for d_3 is 0.051 while the score for d_4 is 0.064. It will then choose d_4 and fail to achieve maximal proportionality although it is possible to do so. Consequently, `PM-2` introduces the parameter λ to

[2] $P(d|q) = 0.3$ is a typical query likelihood score for a document d that is truely relevant to the query q

gain more assurance with regard to the goal that the chosen document *must* be relevant to t^*:

$$d^* \leftarrow \arg\max_{d \in R} \ \lambda \times q_{t^*} \times P(d|t^*) + (1-\lambda) \sum_{t \neq t^*} q_t \times P(d|t) \tag{3.3}$$

It is easy to verify that with $\lambda = 0.8$, d_3 will receive a higher score than d_4.

A second difference between `PM-2` and `PM-1` is that when a document d^* is selected for the current seat, since it is assumed to be relevant to all topics $t \in T$, each topic occupies a certain "portion" of this seat as opposed to a single topic taking up the entire seat as previously. Intuitively, the degree of occupation of the seat is proportional to the normalized relevance to d^*:

$$s_t \leftarrow s_t + \frac{P(d^*|t)}{\sum_{t' \in T} P(d^*|t')}$$

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

`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 topic this seat should go to in order to best maintain the proportionality. Then, it selects the document that, in addition to being relevant to this topic, is relevant to other topics as well. Finally, it updates the "portion" of seats in S occupied by each of the topics $t \in T$ according to how relevant it is to the selected document.

3.3.2.4 Connection to the Novelty-based Approach

Since our techniques rely on a set of query topics, they can be regarded as an *explicit* method. This puts them in the same category as `IA-Select` (Agrawal et al., 2009), the set-based approach (Carterette & Chandar, 2009) and `xQuAD` (Santos et al., 2010a). At a glance, the main difference between our techniques and these is the

optimization objective: proportionality or novelty. In this section, we compare these two approaches in more details.

As surveyed in Chapter 2, the main difference among existing novelty-based techniques is in the way they measure novelty. As a result, we will conduct this comparison using a representative approach from the novelty-based category. We choose `xQuAD` since it has been demonstrated to be among the most effective technique on the TREC corpus that we use for our experiments.

`xQuAD` employs a greedy framework that iteratively selects documents from the input ranking R to form the diverse ranking S. The scoring function is as follows:

$$d^* \leftarrow \arg\max_{d \in R} \ (1-\lambda) \times P(d|q) + \lambda \sum_{t \in T} P(d|t) \times P(t|q) \times \prod_{d_j \in S} (1 - P(d_j|t))$$

While the first component, $P(d|q)$, captures the relevance of the candidate document d to the query q, the second one measures the novelty of d based on the documents that have been selected in the previous iterations. Let $\pi_t = \prod_{d_j \in S}(1 - P(d_j|t))$. π_t can be interpreted as the probability that the topic t has *not* been covered by all documents currently in S. `xQuAD` achieves novelty by promoting documents that are relevant to the topics with lower probability of having been covered. The relative importance of the query topics is captured by $P(t|q)$. Without loss of generality, we can use the topic popularity to represent its importance: $P(t|q) = p_t$. If follows that if we define $\tau_t = p_t \pi_t$, `xQuAD`'s objective function can be rewritten as follows:

$$d^* \leftarrow \arg\max_{d \in R} \ (1-\lambda) \times P(d|q) + \lambda \sum_{t \in T} P(d|t) \times \tau_t$$

The algorithm for `xQuAD` is outlined in Algorithm 6 below.

It can be seen that `PM-2` and `xQuAD` are similar in that they both greedily select one document at a time. At each iteration, `PM-2` computes the quotient q_t for each topic t and `xQuAD` computes τ_t with respect to both the topic popularity and the documents

Algorithm 6 xQuAD

for all slots in the ranked list S **do**
  for all topics $t \in T$ **do**
    $\pi_t = \prod_{d \in S}(1 - P(d|t))$
    $\tau_t = p_t \pi_t$
  end for
  $d^* \leftarrow \text{argmax}_{d \in R} \ (1 - \lambda) \times P(d|q) + \lambda \sum_{t \in T} P(d|t) \times \tau_t$
  $S \leftarrow S \cup \{d^*\}$
  $R \leftarrow R \setminus \{d^*\}$
end for

in S. Although these two quantities are computed differently, they have very similar semantics: they both indicate the importance for this topic t to be satisfied by the current candidate in order to achieve their overall objective. They then score each candidate based on their relevance to all topics weighted by q_t or τ_t to promote multi-topic documents. Intuitively, the two objectives – proportionality and novelty (or redundancy) – are highly correlated. A novel candidate document that is relevant to the topics that have not been well covered by S also makes S more proportional and vice versa. As a result, the way `PM-2` works can also be explained as promoting novelty at every rank using a proportionality-based measure of novelty (since q_t is intended to maximize the proportionality of S). This choice of novelty measure is the first difference between `PM-2` and `xQuAD`.

The second difference is that `PM-2` explicitly distinguishes between the topic with highest q_t and the rest whereas `xQuAD` does not. In fact, the objective function of `xQuAD` has the same form as Formula (3.2). As we have shown in Section 3.3.2.3, this means that it is entirely possible that `xQuAD` would pick a document that is relevant to a handful of already well covered topics over those that provide coverage for fewer topics but includes the most under-represented one. `PM-2` with a large λ value, on the other hand, would do the opposite due to its emphasis on proportionality. Note that we make no claims regarding which behavior will lead to higher user satisfaction but rather point out the difference in the behavior of the two approaches.

Last but not least, xQuAD interpolates the estimated novelty of a document d with its relevance to the query $P(d|q)$ where our formulation of PM-2 does not. In a perfect world where one can accurately determines that $P(d|t) = 1$ for those d that are relevant to t and $P(d|t) = 0$ otherwise, $P(d|q)$ would not be necessary. In practice, however, the estimate of $P(d|t)$ can be erroneous, which can cause the system to return documents on the non-relevant topics. Integrating $P(d|q)$ into the framework can be seen as a a way to make up for this error. In principal, we could incorporate $P(d|q)$ into PM-2 in the same fashion but we choose not to. Instead, we assume the component that estimates the relevance of a document to a query topic, $P(d|t)$, has taken $P(d|q)$ into account. For example, one can use machine learning techniques to estimate $P(d|t)$, in which case $P(d|q)$ can be conveniently used as an additional feature to improve the accuracy of this estimate.

3.4 Summary

We have introduced a new perspective to search result diversification: diversity by proportionality. Instead of quantifying diversity using the amount of novelty in a result list, we consider this list more diverse if the ratio between the number of documents it provides for each query topic matches more closely with the topic popularity distribution. Based on this notion of proportionality, we derived an effectiveness measure which we called Cumulative Proportionality (CPR). We also derived a framework for optimizing proportionality in search results with two instantiations: PM-1 and PM-2. While PM-1 is a naive adaptation of the Sainte-Laguë method, it serves as a basis to implement PM-2, a more practical adaptation that takes into account the fact that a document might be related to multiple topics. In the next chapter, we will compare CPR with the standard diversity measures as well as evaluate our proportionality framework.

CHAPTER 4

FRAMEWORK EVALUATION

In the previous chapter, we have introduced our proportionality approach to search result diversification. This includes an effectiveness measure (CPR) and a proportionality framework with two instantiations (`PM-1` and `PM-2`). In this chapter, we will first compare our proportionality measure with a variety of standard metrics for diversity to study their consistency and more importantly, identify cases where they disagree (Section 4.1).

Recall that diversification has been studied as a re-ranking process. A ranking of documents that are potentially relevant to some of the query topics is first obtained for this query using a relevance-based retrieval model. Diversification techniques, including `PM-1` and `PM-2`, are then used to reorder the documents in this ranking in order to make it more diverse with respect to those topics. We evaluate our framework in the following ways:

- In a controlled environment where we know the set of topics associated with each query and that they are equally popular, can `PM-1` and `PM-2` diversify the results given by the baseline retrieval models effectively?
- How effective are our techniques compared to existing work from both the proportionality and novelty perspectives?
- Do the results depend on which baseline model is used? As more effective baselines are employed, can these techniques still improve diversity?

- How do these techniques perform in the case where some query topics are more popular than others?
- In a practical setting where we do not know the query topics but have to infer them, are these techniques still useful?

In the remaining of this chapter, we will first explain our experimental setup (Section 4.2) and then answer each of these questions with extensive analyses (Section 4.3).

4.1 Comparison of Effectiveness Measures

In order to compare our proportionality measure *CPR* with existing measures for diversity, we collect the runs submitted for the diversity task of TREC Web Track 2009-2012 (Clarke, Craswell, & Soboroff, 2009; Clarke et al., 2010; Clarke, Craswell, Soboroff, & Voorhees, 2011; Clarke & Craswell, 2012). This includes 177 runs. We then compare these runs when evaluated by our *CPR* measure and three standard diversity measures: α-*NDCG* (Clarke et al., 2008), *ERR-IA* (Chapelle et al., 2009) and *NRBP* (Clarke, Kolla, & Vechtomova, 2009). These are the cascade metrics designed to quantify the amount of *novelty* or *redundancy* in a result ranking. For additional analysis, we also compare the correlation between *CPR* and *S-Recall* (Zhai et al., 2003) and *Precision-IA* (Agrawal et al., 2009), both of which are set-based measures from the intent-aware family. *Precision-IA* computes the average precision across all query topics and *S-Recall* is the fraction of query topics for which the search results have at least one relevant document. Details about these measures can be found in Chapter 2. As done in the official TREC evaluations, all measures are computed using the top 20 retrieved documents from each run. Note that the TREC dataset provides a set of ground-truth topics associated with each query without any information about the popularity of these topics. In practice, these topics are often assumed to be equally popular or important.

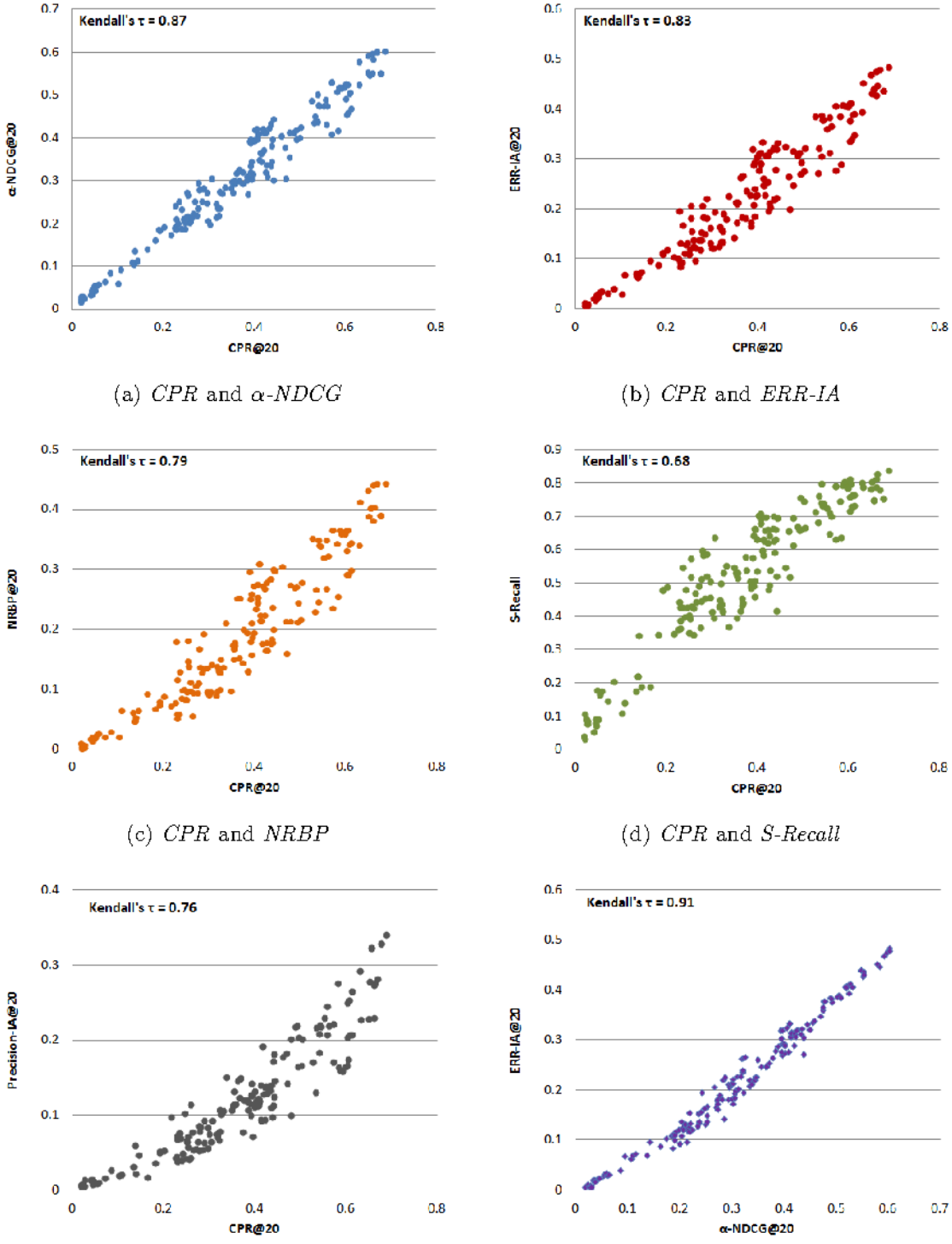
4.1.1 Correlation with Existing Measures

Figure 4.1 shows that *CPR* has quite high correlation with all existing measures. This can be explained as follows. Intuitively, having higher proportionality at every rank is roughly equivalent to having higher novelty, which explains the correlation with the three cascade measures. Furthermore, higher proportionality also indicates having coverage for more topics as well as a considerable number of relevant documents evenly distributed across topics. This explains the correlation with *S-Recall* and *Precision-IA*. Note that, although the correlation between *CPR* and the other measures it quite high, it is lower than the correlation between the two existing measures α-*NDCG* and *ERR-IA*.

To quantify this correlation, we follow Clarke, Craswell, Soboroff, and Ashkan (2011) and use Kendall's τ, which is a well established rank correlation measure. Its values range from -1 to $+1$ with $+1$ representing perfect agreement between two rankings and -1 representing perfect disagreement. Prior work has used $\tau \geq 0.9$ to indicate that two rankings are "equivalent" and $\tau \leq 0.8$ to indicate that there are "noticeable differences" between them (Buckley & Voorhees, 2004).

Figure 4.1 indicates that Kendall's τ value between *CPR* and the cascade measures are quite high (τ value of 0.87, 0.83 and 0.79 for α-*NDCG*, *ERR-IA* and *NRBP* respectively). Intuitively, a ranking with low *S-Recall* or low *Precision-IA* is ineffective. It is not so surprising that different measures have strong agreement on such rankings. It is more interesting to look into the rankings with high *S-Recall* and *Precision-IA* since this is where the difference between measures is most obvious. Figure 4.2 shows the correlation among measures together with their Kendall's τ computed using the 35 runs where *S-Recall* is above 0.5 and *Precision-IA* is above 0.2. Generally, the correlation is substantially lower.

Regarding the two set-based measures, the Kendall's τ value between them and *CPR* are rather low. This is also not surprising because these two measures ignore



(a) *CPR* and α-*NDCG*

(b) *CPR* and *ERR-IA*

(c) *CPR* and *NRBP*

(d) *CPR* and *S-Recall*

(e) *CPR* and *Precision-IA*

(f) α-*NDCG* and *ERR-IA*

Figure 4.1: Correlation between Cumulative Proportionality (*CPR*) and five standard diversity measures: α-*NDCG*, *ERR-IA*, *NRBP* and *S-Recall* and *Precision-IA*.

Table 4.1: Discriminative power of *CPR* and standard diversity measures under the Fisher's randomization test with a significance level of 0.05.

Measure	Discriminative Power
CPR @20	67.12%
α-*NDCG* @20	65.50%
ERR-IA @20	60.26%
NRBP @20	55.87%
S-Recall @20	59.19%
Precision-IA @20	66.07%

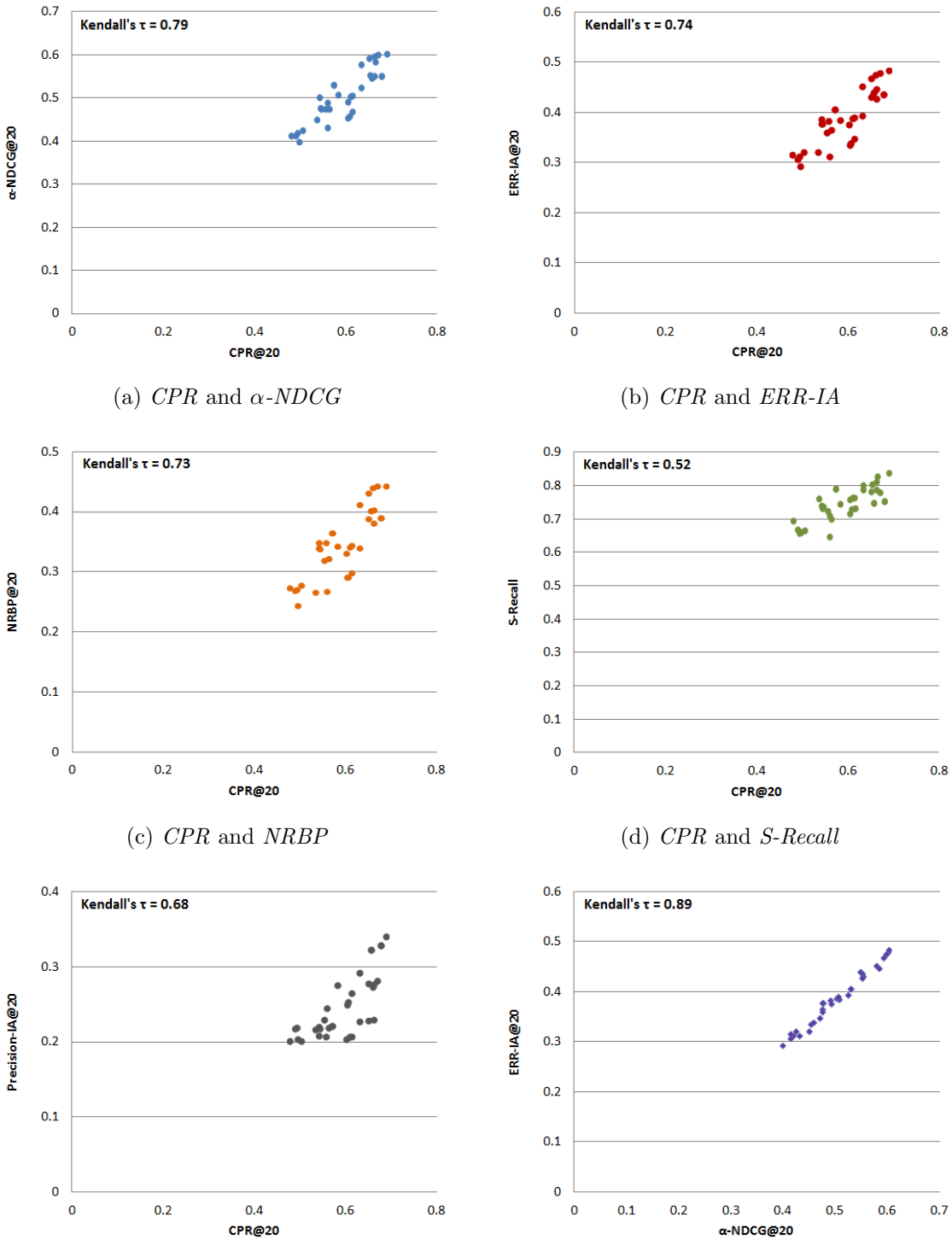
the order among documents. Therefore, in the remaining of this section, we will focus on comparing *CPR* with the cascade measures.

4.1.2 Discriminative Power

Sakai and Song (2011) propose a simple method for assessing the discriminative power of an effectiveness measure which was later used by Clarke, Craswell, Soboroff, and Ashkan (2011) for comparing existing diversity measures. This method performs a significance test between all pairs of retrieval runs under some evaluation measure. The percentage of pairs whose difference is statistically significant at a predefined level is considered the discriminative power of this measure. It expresses the degree to which a metric can detect the differences between systems with high confidence. Sakai and Song (2011) regards "high discriminative power as a necessary condition for a good evaluation metric, not as a sufficient condition". Table 4.1 presents the results for all measures in which we use the Fisher's randomization test with a significance level of 0.05. It shows that *CPR* is the most discriminative metric.

4.1.3 Disagreement with Cascade Measures

CPR and all of the cascade metrics measure diversity as a combination of three factors: topic coverage (as captured by *S-Recall*), the number of documents retrieved for each topics (as captured by *Precision-IA*) and how these documents are ordered. To understand the difference between them, we sort the runs for Web Track 2012 by



(a) *CPR* and *α-NDCG*

(b) *CPR* and *ERR-IA*

(c) *CPR* and *NRBP*

(d) *CPR* and *S-Recall*

(e) *CPR* and *Precision-IA*

(f) *α-NDCG* and *ERR-IA*

Figure 4.2: Correlation between Cumulative Proportionality (*CPR*) and five standard diversity measures: *α-NDCG*, *ERR-IA*, *NRBP* and *S-Recall* and *Precision-IA* measured form the runs where *S-Recall* $\geq$ *0.5* and *Precision-IA* ≥ 0.2

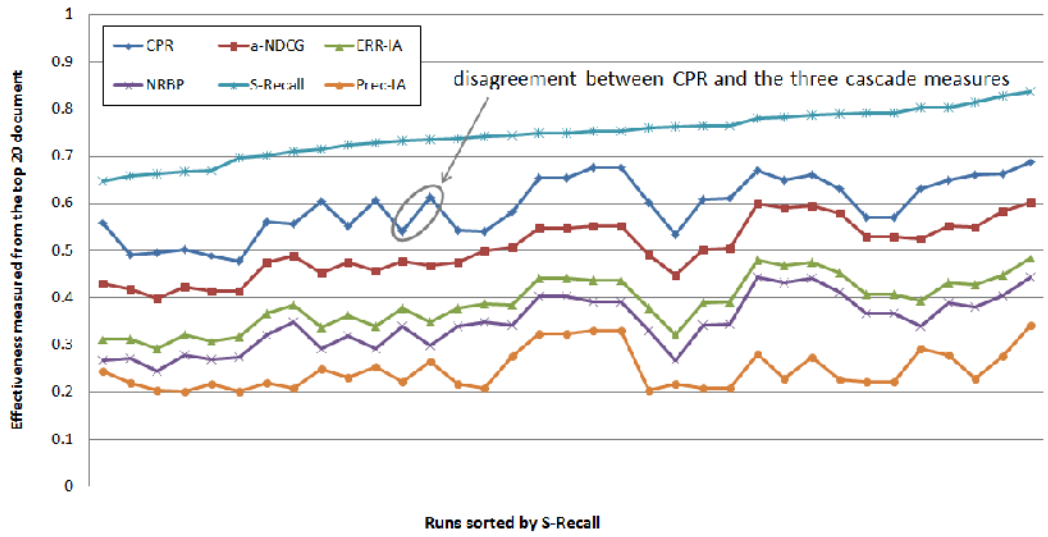


Figure 4.3: TREC 2012 diversity runs evaluated with different measures, sorted by *S-Recall* values.

their *S-Recall* values and plot their effectiveness under these measures in Figure 4.3. Scanning the plot from left to right, we examine the consecutive pairs of runs (note that the run to the right always has higher *S-Recall* due to sorting) where *CPR* disagrees with the cascade measures (i.e. one increases as the others decrease and vice versa). An example of such a run pair is marked in Figure 4.3.

We will focus our analysis on the difference between *CPR* and α-*NDCG*. Our findings will apply to the other two measures as well since α-*NDCG* highly correlates with them. The example pair of runs shown in Figure 4.3 corresponds to an increase in both *S-Recall* and *Precision-IA*. This suggests that the second run covers more topics and has more relevant documents per topic. However, the value of α-*NDCG* decreases. This means that the ordering of the documents in the second run must be rather ineffective. This could mean that the relevant documents are not highly ranked or the documents from the less popular topics are ranked higher than those from the more popular ones. *CPR*, on the other hand, increases. This reveals the difference between *CPR* and the cascade measures: *CPR* puts more emphasis on

Table 4.2: Two pairs of rankings in which one achieves higher α-*NDCG* but lower *CPR* than the other. The two rankings in each pair corresponds to the same query. These results show that *CPR* puts more emphasis on result rankings with high topic coverage and more relevant documents per topic than it does on having the most effective document ordering. α-*NDCG*, as well as other cascade measures, puts more emphasis on the last factor.

Rank	Document	Topic	Document	Topic
	Pair #1: R_1 achieves higher α-*NDCG* but lower *CPR* than R_2			
	R_1		R_2	
1	enwp00-06-18135	t_2, t_4	en0001-76-17061	
2	en0104-87-33374		en0001-76-17975	
3	en0104-87-33373		en0005-32-12986	t_2, t_4
4	en0078-80-03018		en0001-76-08393	
5	en0104-87-33372		en0001-76-17979	
6	enwp00-81-18242	t_2, t_4	en0001-76-17977	
7	en0109-32-34673		en0001-76-17978	
8	en0109-32-34829		enwp00-07-18161	t_2, t_4
9	en0019-27-20755		enwp00-13-18242	t_2, t_3, t_4
10	en0019-27-20762		enwp00-81-18242	t_2, t_4
	Pair #2: R_3 achieves higher α-*NDCG* but lower *CPR* than R_4			
	R_3		R_4	
1	enwp01-80-20189	t_1, t_2, t_3, t_4	en0006-92-01649	t_1, t_2, t_4
2	enwp01-57-19101	t_1, t_2, t_3, t_4	enwp01-66-20688	t_1, t_2, t_3, t_4
3	en0011-07-04255		enwp02-13-21042	t_1, t_2, t_3, t_4
4	enwp02-13-21042	t_1, t_2, t_3, t_4	enwp01-51-18348	t_1, t_2, t_3, t_4
5	en0007-45-09334		enwp01-57-19101	t_1, t_2, t_3, t_4
6	en0004-31-03561	t_2	enwp03-03-01371	t_1, t_2, t_3, t_4
7	enwp03-03-01371	t_1, t_2, t_3, t_4	enwp01-80-20189	t_1, t_2, t_3, t_4
8	en0067-10-30480		enwp01-84-18808	t_1, t_2, t_3, t_4
9	en0006-92-01649	t_1, t_2, t_4	enwp02-29-19446	t_1, t_2, t_3, t_4
10	en0071-39-26704		en0001-83-03341	t_1

having coverage for more topics and more documents for all topics than on having the most effective document ordering, while α-*NDCG* puts more emphasis on the last factor. This can be seen more clearly from the two pairs of rankings (taken from these two runs) presented in Table 4.2. This table shows that *CPR* prefers R_2, which has broader topic coverage and more relevant documents for each topics than R_1. α-*NDCG*, on the other hand, favors R_1, whose first document is a relevant one. Similarly, α-*NDCG* favors R_3 because the first document in this ranking covers one

Table 4.3: Correlation between *CPR* and existing diversity measures in two cases: when all topics are equally popular and when some topics are more popular than others. The 35 runs are those with *S-Recall* $\geq$ *0.5* and *Prec-IA* ≥ 0.2

Correlation	Kendall's τ			
	All 177 runs		35 runs with high *S-Recall* and *Precision-IA*	
	Uniform	Non-uniform	Uniform	Non-uniform
ERR-IA	0.83	0.85	0.74	0.77
NRBP	0.79	0.80	0.73	0.77
Precision-IA	0.76	0.76	0.68	0.67

additional topic compared to the one from R_4. Meanwhile, *CPR* prefers R_4 since it has a much larger number of relevant documents per topic.

4.1.4 Non-uniform Popularity Distribution

To simulate the case where some topics are more popular than others, we follow Sakai and Song (2011) by assuming the j-th topic of a query with n topics has the popularity of $\frac{2^{n-j+1}}{\sum_{k=1}^{n} 2^k}$. We re-evaluate all of the runs using this non-uniform popularity distribution. Table 4.3 provides the Kendall's τ values between *CPR* and the existing measures. We do not consider α-*NDCG* and *S-Recall* since they do not take topic popularity into account. Interestingly, the correlation between *CPR* and *ERR-IA* and *NRBP* is now even higher than before. Table 4.4 presents an example of two rankings where *CPR* and *ERR-IA* disagree under the uniform popularity distribution. While *CPR* prefers R_1, *ERR-IA* prefers R_2. In the non-uniform case, since the topic t_1 has the highest popularity and t_4 has the lowest, both measures come to an agreement that R_1 is more effective. Our explanation for this is that, among the cases where these measures disagree, the main reason for this disagreement, as we presented earlier, is their emphasis on document ordering. As the distribution becomes more skewed, the optimal ordering becomes clearer (i.e., the popular topics should be presented earlier in the ranking), leaving less room for disagreement.

Table 4.4: Two rankings on which *CPR* and *ERR-IA* disagrees in the uniform case: *CPR* prefers R_1 while *ERR-IA* prefers R_2. In the non-uniform case, however, they both agree that R_1 is more effective.

Rank	Document	Topic	Document	Topic
	R_1		R_2	
1	en0009-92-11626		en0007-26-27181	t_4
2	en0009-75-18849		en0002-12-01817	
3	en0009-92-11627	t_1	en0002-12-01843	
4	en0009-03-00665	t_1, t_3, t_4	en0002-12-01846	
5	en0085-43-06520		en0002-12-01815	
6	en0005-47-05526		en0002-33-20829	
7	en0021-64-25478		en0011-14-22813	
8	en0082-31-10635		en0010-45-15999	
9	en0024-67-12684		en0011-39-10739	
10	en0024-67-12691		en0009-92-11627	t_1

4.2 Experimental Setup

In this section, we describe the experimental setup for evaluating our proportionality framework. The results will be presented in the next section.

4.2.1 Query and Retrieval Collection

Our query set consists of 200 queries with relevance judgments from the diversity task of the TREC Web Track 2009-2012 (Clarke, Craswell, & Soboroff, 2009; Clarke et al., 2010; Clarke, Craswell, Soboroff, & Voorhees, 2011; Clarke & Craswell, 2012). Our evaluation is done on the ClueWeb09 Category B retrieval collection, which contains approximately 50 million web pages in English. Details about this corpus have been provided in Chapter 2. Both the queries and the collection are stemmed using the Krovetz stemmer (Krovetz, 1993). In addition, we perform stopword removal only at query time using a small stopword list.

4.2.2 Baseline Retrieval Model

To understand how diversification techniques are affected by the quality of the initial rankings, we consider several models for retrieving them.

4.2.2.1 Query Likelihood

Query likelihood (`QL`) (Ponte & Croft, 1998) is a well established bag-of-words model. It assumes that there are no dependencies between the terms in the user queries and ranks documents based on how likely they are to generate these terms.

Spam filtering is known to be an important component of web retrieval (Bendersky et al., 2010). In addition, documents with too few stopwords are found to have poor readability (Kanungo & Orr, 2009; Ntoulas et al., 2006). Therefore, we incorporate both of these into our baseline ranking. We use the spam filtering technique described by Cormack et al. (2011), which assigns a "spamminess" percentile $S(d)$ to each document d in the collection. Let $\sigma(d)$ be the stopword to non-stopwords ratio in d and $P_{QL}(d|q)$ indicate the score the query likelihood model assigns to the document d. Following Bendersky et al. (2010), the final score of d is given by:

$$P(d|q) = \begin{cases} P_{QL}(d|q) & \text{if } S(d) \geq 60 \text{ and } \sigma(d) \geq 0.1 \\ -\infty & \text{otherwise} \end{cases}$$

4.2.2.2 Sequential Dependence Model

Metzler and Croft (2005) model term dependencies for effective retrieval using a Markov random field (MRF). It models the joint distribution over a document random variable and query term random variables. One can instantiate an MRF model with respect to certain assumptions about the dependency among query terms by defining a corresponding graph structure. Metzler and Croft (2005) provide three such instantiations. In this dissertation, we adopt the *sequential dependence* variant (`SDM`), which assumes dependency between every pair of adjacent query terms. The `SDM` model has been used in many retrieval experiments and has been shown to achieve a good balance between effectiveness and efficiency. Spam filtering and stopword to non-stopword ratio are also integrated in a similar fashion as with query likelihood. In our experiments, model parameters are determined via 5-fold cross validation.

4.2.2.3 Relevance Model

The Relevance Model (`RM`) (Lavrenko & Croft, 2001) is a pseudo-relevance feedback technique developed for the language modeling framework (Croft et al., 2009). It first retrieves a set of potentially relevant documents for a given query, then expands this query with terms that highly co-occur with the query terms within these pseudo-relevant documents. The expanded query is used to retrieve the final ranking. In our experiments, `SDM` is used to retrieve the feedback documents. Similarly, model parameters are determined via 5-fold cross validation, and spam filtering as well as stopword to non-stopword ratio are also incorporated.

4.2.2.4 Learning to Rank with Coordinate Ascent

Unsupervised retrieval models such as query likelihood, sequential dependence and relevance model make use of a very limited number of features such as term frequency, inverse document frequency and document length. The combination of these features is hard-coded into the retrieval model. In contrast, learning to rank (LTR) approaches (Liu, 2009) allow retrieval systems to incorporate hundreds or even thousands of arbitrarily defined features. Most importantly, these approaches automatically learn the most effective combination of these features in the ranking function based on the available training data. As a result, learning to rank approaches have consistently outperformed the traditional models (Liu, 2009).

Learning to rank algorithms can be broadly classified into three approaches: point-wise, pair-wise and list-wise. The point-wise approach attempts to accurately predict the relevance label for individual documents. Pair-wise methods (Freund et al., 2003; Burges et al., 2005, 2006) focus instead on the ability to rank relevant documents higher than the non relevant. List-wise techniques (J. Xu & Li, 2007; Metzler & Croft, 2007; Wu et al., 2010) take the entire ranked list as input and directly optimize

a retrieval measure defined upon this list. Generally, the list-wise approach has been demonstrated to be the most effective.

For our experiments, we use the list-wise technique proposed by Metzler and Croft (2007) to learn a linear combination of features that maximizes Mean Average Precision (MAP). We implement a variety of features used in previous work (Metzler & Croft, 2005; Liu, 2009; McCreadie et al., 2011; Bendersky et al., 2011). Table 4.5 provides an overview of these features. Learning is done using coordinate ascent (`CA`), a well-known technique for unconstrained optimization. It optimizes multivariate objective functions by sequentially doing optimization in one dimension at a time. It cycles through each parameter and optimizes over it while fixing all the others. Note that in this dissertation, we use the term *coordinate ascent*, or `CA`, to refer to this particular ranking technique rather than the general optimization method.

Similar to diversification, LTR is also applied in a two-stage fashion. Firstly, an unsupervised retrieval model is used to retrieve a small set of highly ranked documents from the entire document index. These retrieved documents, together with their human-assigned relevance labels, are then used to train a learning to rank model at the second stage. At run-time, in response to user queries, the unsupervised model is used again to retrieve a small set of highly ranked documents, which are then re-ranked by the trained ranker. Finally, the re-ranked results are presented to the user. In our experiments, we use Relevance Model as the initial retrieval model since it is a recall-oriented approach, which has been shown to be beneficial to LTR systems (Dang et al., 2013).

Recall that the TREC corpus comes with relevance judgments for both the query level (how relevant a document is to a query) and the topic level (whether or not a document is relevant to a particular topic of the query). Our LTR system is trained using the entire set of 200 queries with the associated query level judgments via 5-fold cross validation.

Table 4.5: The set of features we use to trained our learning to rank model.

Feature	**Document Section**
TF, IDF, TF*IDF (min/max/sum/mean/var)	[Body, Anchor, Title, Whole page]
Number of covered query terms	[Body, Anchor, Title, Whole page]
Document length	[Body, Anchor, Title, Whole page]
BM25	Whole page
Query Likelihood (Two-stage/Dirichlet/JM smoothing)	[Body, Anchor, Title, Whole page]
Sequential Dependence (Two-stage/Dirichlet/JM smoothing)	[Body, Anchor, Title, Whole page]
URL Length/Depth	
Number of in-links	
PageRank	
Stopwords fraction/coverage	Whole page
Number of terms/Term entropy	Whole page
Score from M_A^*	Whole page

4.2.3 Diversity Models

We evaluate `PM-2`, the proportionality model we propose for search result diversification. In addition, we will also present results obtained by `PM-1` for comparison. One baseline diversity model for comparison is `MMR` (Carbonell & Goldstein, 1998), 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 (Santos et al., 2010a).

4.2.4 Query Topics

Except for `MMR`, all of the methods under investigation assume the availability of the query topics and their popularity. We first consider the official sub-topics identified by TREC's assessors for each of the queries as the topics. We provide these topics to each system and evaluate their output by how well they cover these topics. This simulates the situation where we know exactly what topics the query has and

provides a controlled environment to study the effectiveness of different diversification approaches. Since TREC's judgment data does not specify the popularity of these topics, we assume that they are equally popular, which is consistent with existing work (Santos et al., 2010a, 2010b, 2011). To simulate the the case where some topics are more popular than others, we follow Sakai and Song (2011) and assume that the i-th topic of a query with n topics has the popularity of $\frac{2^{n-i+1}}{\sum_{j=1}^{n} 2^j}$. The topic popularity is used by both the systems and their evaluation accordingly.

In order to simulate more practical settings in which we do not know but have to guess the topics of the query, we follow Santos et al. (2010a) by adopting suggestions provided by a commercial search engine as topic representations. However, the search engine is unable to provide suggestions for some of the queries in our set. As a result, these experiments are conducted on the subset of 190 queries for which we can obtain topic representations. Each system diversifies the baseline ranking with respect to these suggestions and is evaluated by how well it covers the ground-truth TREC topics. Both topic sets are assumed to have a uniform distribution of popularity.

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

Recall that `PM-1`, `PM-2` and `xQuAD` assume that $P(d|t)$, the relevance of a document d to a query topic t, is available. In our experiments, we treat each topic as a query and use the query likelihood score for each document $P_{QL}(d|t)$ as the estimate of its relevance. This means that for `xQuAD`, we simply substitute $P(d|t)$ with $P_{QL}(d|t)$ in its objective function. `PM-1` and `PM-2`, on the other hand, further assume $P(d|t)$ has taken into account the relevance of this document to the query $P(d|q)$. As a result,

for these two models, we assign $P(d|t) = P_{QL}^{\beta} P(d|q)^{(1-\beta)}$ where $P(d|q)$ is given by the baseline retrieval model which could be `QL`, `SDM`, `RM` or `CA`. Note that we do not do this for `xQuAD` because its framework already takes $P(d|q)$ into account.

4.2.5 Evaluation Metrics

We first report our results using our proportionality metric *CPR*. Since this metric favors our models as they are designed to capture proportionality, we also report the results using several standard metrics that were designed to evaluate existing methods. This includes those used in the official evaluation of the diversity task at TREC (Clarke et al., 2010; Clarke, Craswell, Soboroff, & Voorhees, 2011; Clarke & Craswell, 2012): *α-NDCG*, *ERR-IA*, *NRBP*, *S-Recall* and *Precision-IA*. Unless stated otherwise, all of these measures are computed using the top 20 documents that each system returns. This is done to be consistent with the official TREC evaluation.

4.2.6 Parameter Settings

We use Lemur/Indri [1] to conduct the baseline query likelihood, sequential dependence and relevance model run with the toolkit's default parameter configuration. The learning to rank system is trained using the RankLib package [2] also with its default parameter settings. All of the diversification approaches under evaluation are applied on the top K retrieved documents from each baseline. `MMR`, `xQuAD` and `PM-2` have the parameter λ to tune. We consider λ values in the range $\{0.05, 0.1, 0.15, ..., 1.0\}$. For `PM-1` and `PM-2`, we also consider the same range of values for β. Our 5-fold cross validation enforces complete separation between tuning and testing.

---

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

[2] http://sourceforge.net/p/lemur/wiki/RankLib/

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 use $K = 50$.

4.3 Experimental Results

4.3.1 Diversification with Ground-truth Topics

In this section, we provide the set of ground-truth topics (TREC sub-topics) to all diversification techniques. They diversify the results initially retrieved by a baseline retrieval model for each query with respect to the corresponding ground-truth topics. This provides a controlled environment to verify and compare the effectiveness of different diversification techniques.

4.3.1.1 Proportionality Measure

Table 4.6 shows the *CPR* score at three cut-off points (5, 10 and 20) each technique achieves under different baseline retrieval models (`QL`, `SDM`, `RM`, and `CA`). It also shows the results for both the *uniform* case (where all topics for each query are considered equally popular) and the *non-uniform* case (where some topics are more popular than others). In addition, it provides the Win/Loss ratio – the number queries each system improves and hurts respectively with respect to the initial rankings. The letters b, m, x and p indicate statistically significant differences (p-value < 0.05) to the baseline, `MMR`, `xQuAD` and `PM-1` respectively.

We will first focus on the uniform case. Table 4.6 shows that while the results with `MMR` are generally the same as the initial rankings, `PM-1`, `PM-2` and `xQuAD` consistently outperform the initial rankings across different cut-off points and baseline retrieval techniques. Among the three diversification techniques, the improvement `PM-1` provides is not as large as `PM-2` and `xQuAD` due to its naive assumption that each document is related to only one topic. `PM-2`, by not making this assumption,

consistently outperforms both PM-1 and xQuAD. In addition, PM-2 is also the most robust method since it helps more queries and hurt fewer ones compared to other techniques (except for the case with RM as the baseline in which xQuAD and PM-2 achieve a comparable Win/Loss ratio).

In the non-uniform case, we see a very similar trend but with some differences. Firstly, the improvement each system provides over the baseline is less substantial. This suggests that the fact that some topics are more popular than other makes the task of diversification harder. This is understandable since it requires a stricter order in which the topics have to presented in the search results. Secondly, the difference between PM-2 and xQuAD becomes larger. This is most obvious with *CPR@5*. The average improvement PM-2 provides over xQuAD across four baselines in the uniform case is 1.7%, which increases to 3% in the non-uniform case. In *CPR@10* and *CPR@20*, this improvement is 2.2% and 0.6% respectively in the uniform case and 2.9% and 2.7% in the non-uniform case. Similarly, PM-2 achieves the best Win/Loss ratio.

These results indicate that PM-2 is the most effective technique for maintaining the proportionality of a result ranking, especially in the case where topics are not equally popular. Notice that although xQuAD is designed to optimize for novelty, it also performs well under our proportionality measure. This is because the two objectives are correlated as demonstrated in Section 4.1. Thus, a method that does well on one measure should do well on the other.

It is expected that the results provided by PM-2 would be more proportional than those provided by xQuAD since our technique is designed to directly optimize for proportionality. We will now evaluate the diversification techniques using standard novelty-based measures.

Table 4.6: Performance of all techniques in *CPR* at different cut-off points. Each system diversifies the results provided by the baseline model with respect to the TREC sub-topics. The Win/Loss ratio is with respect to *CPR@20*. The letters b, m, x and p indicate statistically significant differences to the baseline, `MMR`, `xQuAD` and `PM-1` respectively (p-value < 0.05).

			CPR@5	CPR@10	CPR@20	W/L
Uniform	QL	Base	0.4667	0.4981	0.5127	
		MMR	0.4646 (−0.44%)	0.4937 (−0.88%)	0.5077_{b} (−0.98%)	53/86
		xQuAD	$0.5832_{b,m}$ (+24.96%)	$0.6008_{b,m}$ (+20.63%)	$0.5993_{b,m}$ (+16.9%)	117/57
		PM-1	$0.5375^{x}_{b,m}$ (+15.17%)	$0.5555^{x}_{b,m}$ (+11.54%)	$0.5539^{x}_{b,m}$ (+8.03%)	99/75
		PM-2	$\mathbf{0.597}^{p}_{b,m}$ (+27.91%)	$\mathbf{0.612}^{x,p}_{b,m}$ (+22.87%)	$\mathbf{0.606}^{p}_{b,m}$ (+18.2%)	**123/50**
	SDM	Base	0.5091	0.5349	0.5452	
		MMR	0.4949_{b} (−2.78%)	0.5233_{b} (−2.17%)	0.5382_{b} (−1.28%)	39/110
		xQuAD	$0.5922_{b,m}$ (+16.33%)	$0.6095_{b,m}$ (+13.95%)	$0.608_{b,m}$ (+11.53%)	111/66
		PM-1	$0.2955^{x}_{b,m}$ (−41.96%)	$0.3143^{x}_{b,m}$ (−41.23%)	$0.3554^{x}_{b,m}$ (−34.81%)	26/151
		PM-2	$\mathbf{0.602}^{p}_{b,m}$ (+18.25%)	$\mathbf{0.6176}^{p}_{b,m}$ (+15.46%)	$\mathbf{0.6141}^{x,p}_{b,m}$ (+12.65%)	**118/58**
	RM	Base	0.5138	0.5395	0.5489	
		MMR	0.514 (+0.05%)	0.537 (−0.46%)	0.5469 (−0.36%)	59/85
		xQuAD	$0.5826_{b,m}$ (+13.4%)	$0.602_{b,m}$ (+11.57%)	$0.5982_{b,m}$ (+8.98%)	**100/73**
		PM-1	$0.3215^{x}_{b,m}$ (−37.43%)	$0.3383^{x}_{b,m}$ (−37.3%)	$0.387^{x}_{b,m}$ (−29.49%)	30/144
		PM-2	$\mathbf{0.5943}^{p}_{b,m}$ (+15.67%)	$\mathbf{0.6069}^{p}_{b,m}$ (+12.48%)	$\mathbf{0.6019}^{p}_{b,m}$ (+9.65%)	99/74
	CA	Base	0.5608	0.5814	0.5881	
		MMR	0.5615 (+0.12%)	0.5818 (+0.07%)	0.5882 (+0.02%)	38/48
		xQuAD	$0.6171_{b,m}$ (+10.04%)	$\mathbf{0.6316}_{b,m}$ (+8.63%)	$0.6274_{b,m}$ (+6.67%)	95/81
		PM-1	$0.3505^{x}_{b,m}$ (−37.51%)	$0.3638^{x}_{b,m}$ (−37.42%)	$0.4168^{x}_{b,m}$ (−29.13%)	30/142
		PM-2	$\mathbf{0.6224}^{p}_{b,m}$ (+10.97%)	$\mathbf{0.6316}^{p}_{b,m}$ (+8.63%)	$\mathbf{0.6278}^{p}_{b,m}$ (+6.75%)	**104/74**
Non-Uniform	QL	Base	0.4618	0.4912	0.505	
		MMR	0.4591 (−0.59%)	0.4856_{b} (−1.14%)	0.4991_{b} (−1.16%)	49/91
		xQuAD	$0.5628_{b,m}$ (+21.87%)	$0.5759_{b,m}$ (+17.23%)	$0.5729_{b,m}$ (+13.46%)	104/72
		PM-1	$0.5463_{b,m}$ (+18.3%)	$0.5546_{b,m}$ (+12.9%)	$0.5469^{x}_{b,m}$ (+8.31%)	101/73
		PM-2	$\mathbf{0.5867}^{x,p}_{b,m}$ (+27.05%)	$\mathbf{0.5975}^{x,p}_{b,m}$ (+21.64%)	$\mathbf{0.5902}^{x,p}_{b,m}$ (+16.89%)	**114/64**
	SDM	Base	0.5081	0.5313	0.5403	
		MMR	0.4937_{b} (−2.82%)	0.5197_{b} (−2.17%)	0.5331_{b} (−1.32%)	41/108
		xQuAD	$0.571_{b,m}$ (+12.38%)	$0.5793_{b,m}$ (+9.05%)	$0.5794_{b,m}$ (+7.25%)	108/70
		PM-1	0.5432_{m} (+6.92%)	0.5547^{x}_{m} (+4.42%)	0.5536^{x} (+2.46%)	94/84
		PM-2	$\mathbf{0.5817}^{p}_{b,m}$ (+14.48%)	$\mathbf{0.5975}^{p}_{b,m}$ (+12.47%)	$\mathbf{0.5984}^{x,p}_{b,m}$ (+10.75%)	**116/62**
	RM	Base	0.5152	0.5396	0.5463	
		MMR	0.5183 (+0.6%)	0.5386 (−0.19%)	0.546 (−0.05%)	67/73
		xQuAD	$0.5713_{b,m}$ (+10.89%)	$0.5826_{b,m}$ (+7.97%)	$0.5766_{b,m}$ (+5.56%)	91/83
		PM-1	$0.5642_{b,m}$ (+9.51%)	0.5718_{m} (+5.96%)	0.5672 (+3.83%)	85/89
		PM-2	$\mathbf{0.586}_{b,m}$ (+13.75%)	$\mathbf{0.5966}^{p}_{b,m}$ (+10.55%)	$\mathbf{0.5912}^{x,p}_{b,m}$ (+8.23%)	**104/63**
	CA	Base	0.5601	0.5809	0.586	
		MMR	0.558 (−0.38%)	0.5798 (−0.19%)	0.5852 (−0.15%)	51/67
		xQuAD	$0.6095_{b,m}$ (+8.82%)	$0.6198_{b,m}$ (+6.7%)	$0.6124_{b,m}$ (+4.5%)	88/85
		PM-1	0.6012_{m} (+7.34%)	0.6079 (+4.65%)	0.6007 (+2.5%)	83/96
		PM-2	$\mathbf{0.6305}^{x,p}_{b,m}$ (+12.57%)	$\mathbf{0.6353}^{p}_{b,m}$ (+9.36%)	$\mathbf{0.6254}^{p}_{b,m}$ (+6.72%)	**99/75**

4.3.1.2 Novelty-based Measures

Table 4.7 compares all techniques using standard novelty-based measures. Interestingly, we see a very similar trend as in the previous case with a proportionality measure. `MMR` is the least effective method due to its lack of awareness of the query topics. `PM-2`, on the other hand, outperforms all other methods in almost all metrics with statistically significant improvement in many cases. This is consistent across all four baseline models. In addition, `PM-2` is the most robust technique with the most effective Win/Loss ratio.

Regarding uniform and non-uniform distribution of popularity, with respect to *Precision-IA*, the improvement `PM-2` provides over `xQuAD` is also substantially higher in the non-uniform case (3.2% vs. 0.3%). This is, in fact, consistent with what we observed with *CPR*. This can be explained by the fact that larger improvement in *CPR* indicates that the result rankings have more relevant documents for the more popular topics, which results in higher *Precision-IA*.

Examining *ERR-IA* and *NRBP*, however, we observe the opposite effect. `PM-2` outperforms `xQuAD` by an average of 2.3% and 2.6% in *ERR-IA* and *NRBP* respectively in the uniform case (across four baselines). These numbers go down to 2.1% and 2.3% when the distribution is not uniform. Combined with the fact that *Precision-IA* is higher in the non-uniform case, this provides more evidence that *ERR-IA* puts more emphasis on the document ranking, whereas *CPR* is concerned more with topic coverage.

Given that `PM-2` optimizes proportionality, it is very encouraging to see that `PM-2` manages to outperform `xQuAD` in both cases using these measures of novelty. It confirms the effectiveness of `PM-2`: this technique retrieves result rankings that are not only more proportional but also less redundant compared to `xQuAD`, which is designed to minimize redundancy.

Table 4.7: Performance of all techniques in several standard redundancy-based measures. Each system diversifies the results provided by the baseline model with respect to the TREC sub-topics. The Win/Loss ratio is with respect to α-NDCG. The letters b, m, x and p indicate statistically significant differences to the baseline, `MMR`, `xQuAD` and `PM-1` respectively (p-value < 0.05).

			α-NDCG	ERR	NRBP	P-IA	S-Recall	W/L
Uniform	QL	Base	0.4156	0.3054	0.2679	0.1897	0.6215	
		MMR	0.4109_{b}	0.3018_{b}	0.2641	0.1882	0.614	55/85
		xQuAD	$0.4936_{b,m}$	$0.3743_{b,m}$	$0.3395_{b,m}$	$\mathbf{0.2208}_{b,m}$	$0.669_{b,m}$	121/53
		PM-1	$0.4567^{x}_{b,m}$	0.3324^{x}_{m}	0.2906^{x}	0.1893^{x}	$0.6702_{b,m}$	93/81
		PM-2	$\mathbf{0.5011}^{p}_{b,m}$	$\mathbf{0.3828}^{p}_{b,m}$	$\mathbf{0.3478}^{p}_{b,m}$	$0.2201^{p}_{b,m}$	$\mathbf{0.6745}_{b,m}$	**121/51**
	SDM	Base	0.4393	0.329	0.2933	0.214	0.6204	
		MMR	0.4343_{b}	0.3226_{b}	0.2847_{b}	0.2106_{b}	0.6272	43/105
		xQuAD	$0.5028_{b,m}$	$0.3835_{b,m}$	$0.3491_{b,m}$	$\mathbf{0.2337}_{b,m}$	$0.6746_{b,m}$	108/69
		PM-1	0.4604^{x}_{m}	0.3342^{x}	0.2911^{x}	$0.1995^{x}_{b,m}$	$0.6716_{b,m}$	90/88
		PM-2	$\mathbf{0.5094}^{p}_{b,m}$	$\mathbf{0.3878}^{p}_{b,m}$	$\mathbf{0.3525}^{p}_{b,m}$	$0.2329^{p}_{b,m}$	$\mathbf{0.6787}_{b,m}$	**113/63**
	RM	Base	0.4406	0.3413	0.3106	0.221	0.5978	
		MMR	0.4407	0.3399	0.3083	0.2206	0.6035	61/79
		xQuAD	$0.49_{b,m}$	$0.3767_{b,m}$	$0.3438_{b,m}$	0.2283	$0.6558_{b,m}$	104/69
		PM-1	$0.4659^{x}_{b,m}$	0.3464^{x}	0.3073^{x}	0.2124^{x}	$0.6581_{b,m}$	80/92
		PM-2	$\mathbf{0.5023}^{x,p}_{b,m}$	$\mathbf{0.3878}^{p}_{b,m}$	$\mathbf{0.3563}^{p}_{b,m}$	$\mathbf{0.2299}^{p}$	$\mathbf{0.6669}^{x}_{b,m}$	**116/57**
	CA	Base	0.4824	0.3719	0.3406	0.2457	0.6539	
		MMR	0.4826	0.372	0.3407	0.2449_{b}	0.6552	41/46
		xQuAD	$0.516_{b,m}$	0.3979	0.3647	0.2448	0.6779_{b}	101/75
		PM-1	0.4917^{x}	0.3706^{x}	0.3319^{x}	$0.225^{x}_{b,m}$	0.6766_{b}	88/91
		PM-2	$\mathbf{0.5257}^{p}_{b,m}$	$\mathbf{0.4089}^{p}_{b,m}$	$\mathbf{0.3763}^{p}_{b,m}$	$\mathbf{0.2472}^{p}$	$\mathbf{0.6851}_{b,m}$	**111/68**
Non-Uniform	QL	Base	0.4156	0.3329	0.0873	0.1897	0.6215	
		MMR	0.4109_{b}	0.3288_{b}	0.086	0.1882	0.614	55/85
		xQuAD	$0.464_{b,m}$	$0.4057_{b,m}$	$0.1129_{b,m}$	$0.2123_{b,m}$	$0.656_{b,m}$	104/72
		PM-1	$0.4562_{b,m}$	$0.3841^{x}_{b,m}$	$0.105^{x}_{b,m}$	0.1894^{x}	$0.6653_{b,m}$	96/79
		PM-2	$\mathbf{0.4903}^{x,p}_{b,m}$	$\mathbf{0.4183}^{p}_{b,m}$	$\mathbf{0.1166}^{p}_{b,m}$	$\mathbf{0.2179}^{x,p}_{b,m}$	$\mathbf{0.679}^{x}_{b,m}$	**114/64**
	SDM	Base	0.4393	0.362	0.0969	0.214	0.6204	
		MMR	0.4343_{b}	0.355_{b}	0.094_{b}	0.2106_{b}	0.6272	43/105
		xQuAD	$0.4802_{b,m}$	$\mathbf{0.4128}_{b,m}$	$\mathbf{0.1159}_{b,m}$	$0.2271_{b,m}$	$0.6703_{b,m}$	97/81
		PM-1	0.4612^{x}_{m}	0.3846^{x}_{m}	0.1049^{x}	0.1993^{x}_{b}	$0.6742_{b,m}$	89/89
		PM-2	$\mathbf{0.4954}^{x,p}_{b,m}$	$0.4085^{p}_{b,m}$	$0.1144^{p}_{b,m}$	$\mathbf{0.2351}^{x,p}_{b,m}$	$\mathbf{0.6779}_{b,m}$	**110/68**
	RM	Base	0.4406	0.375	0.1027	0.221	0.5978	
		MMR	0.4407	0.3734	0.1014	0.2206	0.6035	61/79
		xQuAD	$0.4717_{b,m}$	$0.4121_{b,m}$	0.1156_{m}	0.2231	$0.6525_{b,m}$	92/82
		PM-1	0.4646	0.3991	0.1104	$0.2079^{x}_{b,m}$	$0.6544_{b,m}$	86/88
		PM-2	$\mathbf{0.4937}^{x,p}_{b,m}$	$\mathbf{0.4177}_{b,m}$	$\mathbf{0.1178}_{b,m}$	$\mathbf{0.2297}^{x,p}$	$\mathbf{0.6646}_{b,m}$	**109/64**
	CA	Base	0.4824	0.4077	0.1138	0.2457	0.6539	
		MMR	0.4826	0.408	0.114	0.2449_{b}	0.6552	41/46
		xQuAD	0.4867	0.4226	0.1175	0.2373_{b}	0.6619	92/81
		PM-1	0.4922	0.4259	0.1182	$0.2269_{b,m}$	$\mathbf{0.6793}_{b,m}$	90/90
		PM-2	$\mathbf{0.517}^{x,p}_{b,m}$	$\mathbf{0.443}_{b,m}$	**0.1237**	$\mathbf{0.246}^{x,p}$	0.6723	**106/72**

4.3.1.3 Comparative Analysis

Recall that there are three main differences between PM-2 and xQuAD:

(1) Although PM-2 is designed to optimize for proportionality, it can be regarded as a novelty-based approach with a different approach to measuring document novelty. xQuAD estimates the novelty of a document based on how well it satisfies the topics that have low probability of having been satisfied, whereas PM-2 estimates novelty based on how proportional the the result set becomes if this document is selected.

(2) PM-2 uses the parameter λ to put more emphasis on the fact that the document selected at each iteration must be relevant to the most under-represented topic, which is the key to maintain proportionality. xQuAD, on the other hand, does not distinguish between the most under-represented topic and the others.

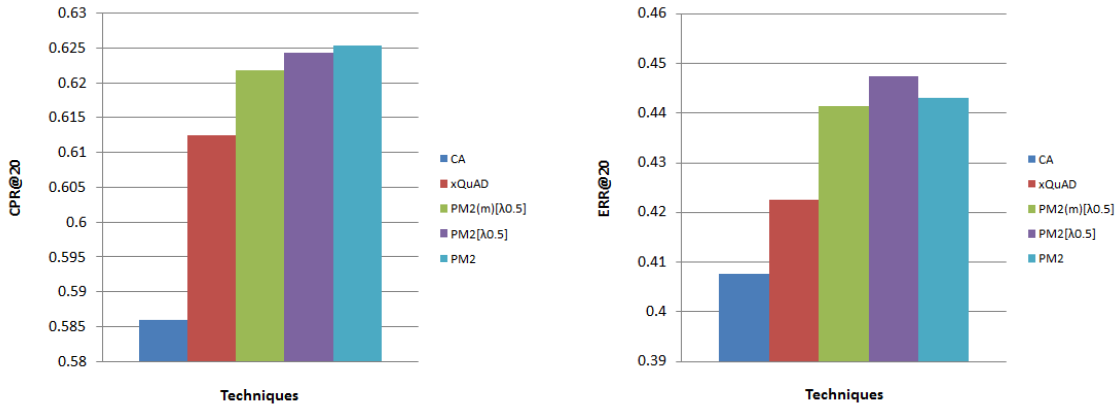
(3) The methods incorporate the relevance of the document to the query into their framework differently. While xQuAD scores a candidate document by a linear combination of its relevance and its novelty, PM-2 integrates this quantity into the component that estimates novelty.

To understand how each of these factors affects their performance, we present experiments with some variants of PM-2 as follows. We conduct a PM-2 run in which we set $\lambda = 0.5$ instead of tuning it via cross-validation as we have done earlier. We will refer to this run as PM-2[$\lambda_{0.5}$] and the cross-validation run as just PM-2 as before. The objective function of PM-2[$\lambda_{0.5}$] can be rewritten as follows:

$$score(d) = \ 0.5 \times q_{t^*} \times P(d|t^*) + 0.5 \sum_{t \neq t^*} q_t \times P(d|t)$$

which is equivalent to:

$$score(d) = \ \sum_{t \in T} q_t \times P(d|t)$$



(a) *CPR* (b) *ERR-IA*

Figure 4.4: Comparison among the baseline `CA`, `xQuAD`, `PM-2` and two of its variants: `PM-2`$[\lambda_{0.5}]$ and `PM-2`$^{(m)}[\lambda_{0.5}]$.

We then linearly combine this objective with the relevance of the document to the query $P(d|q)$:

$$score^{(m)}(d) = (1-\lambda) \times P(d|q) + \lambda \sum_{t \in T} q_t \times P(d|t) \tag{4.1}$$

We will refer to the `PM-2` variant that uses the objective function given by Formula (4.1) above as `PM-2`$^{(m)}[\lambda_{0.5}]$.

Since the only difference between `PM-2`$^{(m)}[\lambda_{0.5}]$ and `xQuAD` is the novelty estimate (we have eliminated factor (2) and (3)), comparing their performance will tell us whether (1) provides `PM-2` with any advantage over `xQuAD`. Additionally, comparing `PM-2`$^{(m)}[\lambda_{0.5}]$ to `PM-2`$[\lambda_{0.5}]$ will reveal the impact of (3) because the only difference between them is in the way $P(d|q)$ is integrated in to their objective function. Finally, the difference between `PM-2`$[\lambda_{0.5}]$ and `PM-2` relates to the effect of factor (2).

Figure 4.4 (a) and (b) show that `PM-2`$^{(m)}[\lambda_{0.5}]$ outperforms `xQuAD` in both *CPR* and *ERR-IA*, which indicates that (1) has very positive impact on `PM-2`'s superiority. In other words, choosing a document based on how proportional the result set becomes is more effective than selecting one based on how well it covers the topics with low

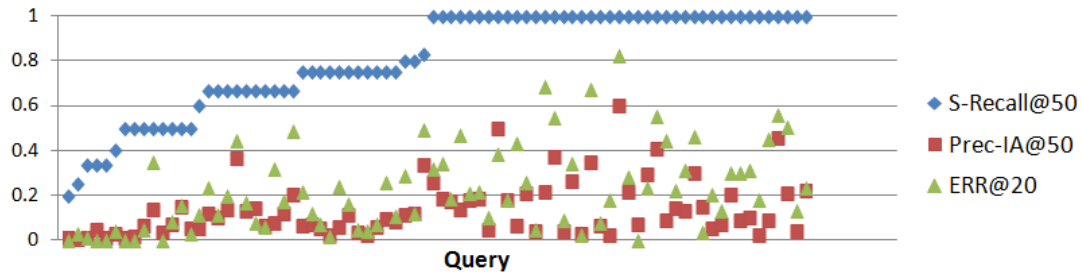
probability of having been covered. Not only does this make the final ranking more proportional, it also reduces the amount of redundancy.

Furthermore, the fact that PM-2[$\lambda_{0.5}$] beats PM-2$^{(m)}$[$\lambda_{0.5}$] in both measures means that factor (3) indeed provides additional benefits to PM-2. This suggests to some extent that it is better to incorporate $P(d|q)$ directly into the estimation of $P(d|t)$.

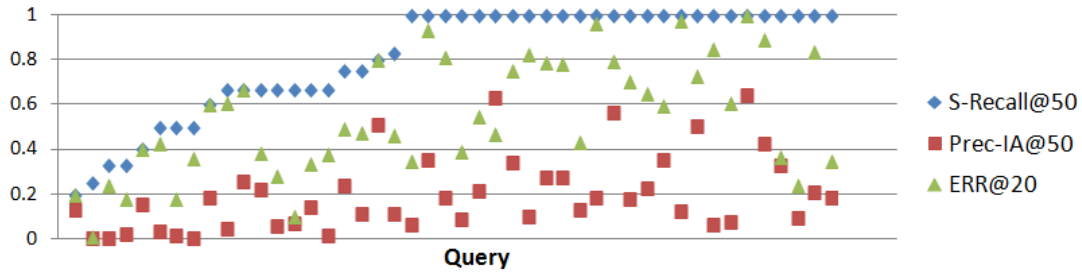
Regarding factor (2), however, we observe mixed results. While tuning λ improves *CPR* somewhat (PM-2 $>$ PM-2[$\lambda_{0.5}$] in Figure 4.4 (a)), it hurts *ERR-IA* (PM-2 $<$ PM-2[$\lambda_{0.5}$] in Figure 4.4 (b)). The value for λ selected via cross validation is $\lambda^* \approx 0.56$, which indicates that PM-2 puts considerable emphasis on the fact that the document selected at each iteration should cover the most under-represented topic at that point. The fact that *ERR-IA* is better with $\lambda = 0.5$ suggests that this measure favors documents that are relevant to multiple relatively well-covered topics than those that are relevant only to the most under-represented topic.

4.3.1.4 Failure Analysis

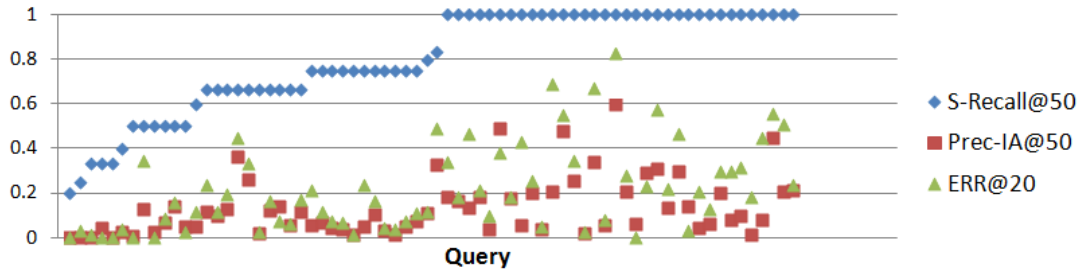
The effectiveness of each model depends on two factors: the quality of the initial retrieved set of documents and the model's power to select a diverse ranking of documents from that pool. The latter factor is apparently affected by the accuracy of the $P(d|t)$ estimate. To understand which of these factors are responsible for the cases that PM-2 fails to provide improvement, we look into the queries which PM-2 improves and hurts by at least 10% *ERR-IA@20*. For each of these queries, we measure *S-Recall@50* and *Prec-IA@50* of the baseline ranking. This indicates how much room there is for PM-2 to improve. We sort all queries in the order of increasing *S-Recall@50* and plot their *Prec-IA@50* as well as *ERR-IA@20*, which reflects the diversity effectiveness of this baseline ranking. Figure 4.5 shows this plot for the case where CA is used as the baseline. The results with other baselines are similar.



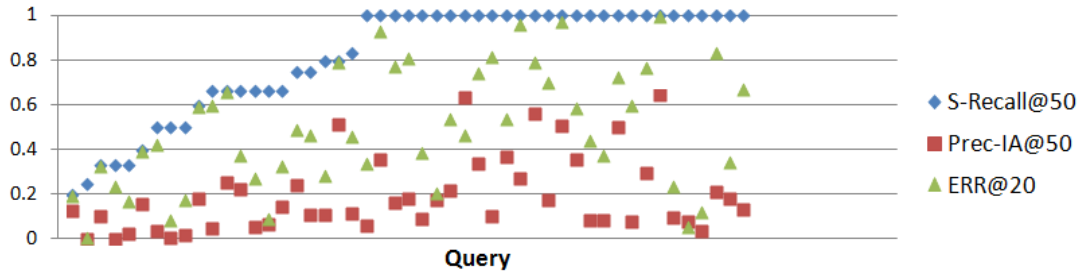
(a) 80 queries improved by `PM-2`



(b) 46 queries hurt by `PM-2`



(c) 70 queries improved by `xQuAD`



(d) 49 queries hurt by `xQuAD`

Figure 4.5: *S-Recall@50* and *Prec-IA@50* evaluated using top 50 documents in the baseline ranking provided by `CA` for each query. *ERR-IA@20* is the diversity effectiveness of this baseline ranking.

Ideally, a diversification technique should improve the baseline rankings that are ineffective while retaining those that are effective, instead of hurting them. Unfortunately, Figure 4.5 (a) and (b) show that this is not the case with PM-2. While it helps a large number of queries, it also hurts a handful of those for which CA retrieves good result rankings (i.e. high *ERR-IA@20*). Since it is not necessary to have high *S-Recall@50* and *Prec-IA@50* in order to keep an effective baseline ranking intact, we argue that the reason that PM-2 fails is that query likelihood is not sufficiently accurate for estimating how relevant a document is to each query topic. This applies to xQuAD as well (see Figure 4.5 (c) and (d)). We believe that improving this estimate will make PM-2 more competitive.

4.3.2 Diversification with Related Queries as Topics

In this section, we obtain from a commercial search engine a set of related queries for each of queries in our dataset. Each technique then diversifies the search results provided by a baseline model using these related queries as topics. The techniques are evaluated based on how well they cover the TREC sub-topics, which we consider ground-truth. Table 4.8 and Table 4.9 compare all techniques being studied using *CPR* and the standard diversity measures respectively.

Overall, the performance of all techniques is lower than in the previous case due to the mis-alignment between the generated topics and the ground-truth. Our findings, on the other hand, are the same. PM-2 outperforms all other techniques in almost all measures. This is also consistent across all baseline retrieval models. This confirms the feasibility of diversification in general and the PM-2 technique in particular. In practice, we often have to predict the topics underlying a user query and this predicted set is not going be completely aligned with the user's interests. Even with this mis-alignment, PM-2 can still outperform other techniques in terms of making the

Table 4.8: Performance of all techniques in *CPR* at different cut-off points. Each system diversifies the results provided by the baseline model with respect to the related queries obtained from a commercial search engines. The Win/Loss ratio is with respect to CPR@20. The letters b, m, x and p indicate statistically significant differences to the baseline, `MMR`, `xQuAD` and `PM-1` respectively (p-value < 0.05).

			CPR@5	CPR@10	CPR@20	W/L
Uniform	QL	Base	0.4642	0.4939	0.5076	
		MMR	0.4627 (−0.33%)	0.4903 (−0.73%)	0.5031_{b} (−0.9%)	53/81
		xQuAD	$\mathbf{0.5105}_{b,m}$ (+9.98%)	$0.5428_{b,m}$ (+9.9%)	$0.5486_{b,m}$ (+8.08%)	93/70
		PM-1	0.4215^{x} (−9.18%)	0.4639^{x} (−6.07%)	0.4891^{x} (−3.64%)	74/98
		PM-2	$0.5055^{p}_{b,m}$ (+8.91%)	$\mathbf{0.5446}^{p}_{b,m}$ (+10.26%)	$\mathbf{0.5537}^{p}_{b,m}$ (+9.07%)	**99/68**
	SDM	Base	0.5036	0.5302	0.5397	
		MMR	0.4895_{b} (−2.82%)	0.5188_{b} (−2.15%)	0.5331_{b} (−1.22%)	38/105
		xQuAD	0.5265_{m} (+4.54%)	$0.5586_{b,m}$ (+5.36%)	$0.5659_{b,m}$ (+4.84%)	93/72
		PM-1	$0.2996^{x}_{b,m}$ (−40.51%)	$0.3137^{x}_{b,m}$ (−40.83%)	$0.3553^{x}_{b,m}$ (−34.17%)	18/149
		PM-2	$\mathbf{0.5294}^{p}_{b,m}$ (+5.12%)	$\mathbf{0.5635}^{p}_{b,m}$ (+6.29%)	$\mathbf{0.5733}^{p}_{b,m}$ (+6.22%)	**100/60**
	RM	Base	0.5069	0.5334	0.5424	
		MMR	0.5079 (+0.19%)	0.5309 (−0.45%)	0.5404 (−0.35%)	57/81
		xQuAD	0.5307 (+4.69%)	0.557_{m} (+4.44%)	0.5641_{m} (+4.0%)	88/74
		PM-1	$0.3148^{x}_{b,m}$ (−37.91%)	$0.3295^{x}_{b,m}$ (−38.22%)	$0.3685^{x}_{b,m}$ (−32.06%)	19/139
		PM-2	$\mathbf{0.5347}^{p}$ (+5.48%)	$\mathbf{0.5632}^{p}_{b,m}$ (+5.6%)	$\mathbf{0.5713}^{p}_{b,m}$ (+5.34%)	90/77
	CA	Base	0.5535	0.5743	0.5813	
		MMR	0.5542 (+0.12%)	0.5749 (+0.09%)	0.5814 (+0.03%)	38/46
		xQuAD	0.5401 (−2.42%)	0.5689 (−0.95%)	0.5799 (−0.23%)	80/89
		PM-1	$0.3513^{x}_{b,m}$ (−36.53%)	$0.3648^{x}_{b,m}$ (−36.49%)	$0.4151^{x}_{b,m}$ (−28.58%)	25/140
		PM-2	$\mathbf{0.56}^{x,p}$ (+1.18%)	$\mathbf{0.5817}^{p}$ (+1.29%)	$\mathbf{0.5863}^{p}$ (+0.87%)	**85/84**

results returned by traditional IR models more effective of satisfying the diverse user information needs.

4.4 Summary

In this chapter, we have evaluated our proportionality approach to search results diversification. We demonstrated that, although our Cumulative Proportionality measure (*CPR*) correlates rather well with existing novelty-based measures, there are differences. Generally, the diversity effectiveness of a document ranking is a combination of three factors: topic coverage, the number of relevant documents for each topic and the order in which these documents are presented. Our analyses using 177 runs submitted to TREC Web Track 2009-2012 showed that the disagreement between our measure and the existing ones is in the marginal cases where a ranking is good in one criteria but bad in others. In particular, we found that *CPR* puts

Table 4.9: Performance of all techniques in several standard redundancy-based measures. Each system diversifies the results provided by the baseline model with respect to the related queries obtained from a commercial search engines. The Win/Loss ratio is with respect to α-NDCG. The letters b, m, x and p indicate statistically significant differences to the baseline, `MMR`, `xQuAD` and `PM-1` respectively (p-value < 0.05).

			α-NDCG	ERR	NRBP	P-IA	S-Recall	W/L
Uniform	QL	Base	0.4111	0.3025	0.2653	0.1844	0.6159	
		MMR	0.4064_{b}	0.2991	0.2618	0.1825	0.6082	53/82
		xQuAD	$0.4402_{b,m}$	0.3256_{m}	**0.2861**	$0.1976_{b,m}$	$0.6389_{b,m}$	91/73
		PM-1	0.391^{x}	$0.2648^{x}_{b,m}$	$0.2151^{x}_{b,m}$	$0.1612^{x}_{b,m}$	$0.6482_{b,m}$	82/90
		PM-2	$\mathbf{0.4455}^{p}_{b,m}$	$\mathbf{0.3267}^{p}_{b,m}$	0.2857^{p}_{m}	$\mathbf{0.1982}^{p}_{b,m}$	$\mathbf{0.6509}^{x}_{b,m}$	**100/66**
	SDM	Base	0.4329	0.3233	0.2874	0.2077	0.6161	
		MMR	0.4279_{b}	0.3168_{b}	0.2786_{b}	0.2046_{b}	0.6231	43/100
		xQuAD	0.4483_{m}	0.329	0.2884	**0.2163**	$\mathbf{0.642}_{b,m}$	92/69
		PM-1	0.4329_{m}	0.3233_{m}	0.2874_{m}	0.2077_{m}	0.6161^{x}	0/0
		PM-2	$\mathbf{0.4587}^{x,p}_{b,m}$	$\mathbf{0.343}^{x,p}_{b,m}$	$\mathbf{0.3052}^{x,p}_{b,m}$	0.2128	$0.6417^{p}_{b,m}$	**97/64**
	RM	Base	0.4346	0.3368	0.3062	0.2128	0.5944	
		MMR	0.4354	0.3357	0.3042	0.2124	0.602	59/75
		xQuAD	0.45	0.3404	0.3053	0.2146	$0.6302_{b,m}$	**90/67**
		PM-1	$0.4153^{x}_{b,m}$	$0.3071^{x}_{b,m}$	$0.2673^{x}_{b,m}$	$0.1957^{x}_{b,m}$	0.6084^{x}	22/47
		PM-2	$\mathbf{0.4549}^{p}_{b,m}$	$\mathbf{0.3428}^{p}$	$\mathbf{0.3071}^{p}$	$\mathbf{0.217}^{p}$	$\mathbf{0.6343}^{p}_{b,m}$	89/75
	CA	Base	0.4753	0.3653	0.3336	0.2382	0.6508	
		MMR	0.4755	0.3655	0.3337	0.2373_{b}	0.6521	40/45
		xQuAD	0.478	0.371	0.3409	$0.2321_{b,m}$	0.6453	98/65
		PM-1	0.4779	0.373	0.3439	$0.2314_{b,m}$	0.6416	98/66
		PM-2	**0.4849**	**0.3772**	**0.3469**	0.2342	**0.6548**	**102/63**

more emphasis on the fact that a ranking should have broad topic coverage and more relevant documents per topic whereas existing metrics gives higher reward to those with lower topic coverage but present relevant documents at slightly higher positions.

Our results have also shown that our diversification method `PM-2` consistently provides significant improvement over four standard relevance-based retrieval models. Additionally, with both manually and automatically generated query topics, `PM-2` outperforms the top performing redundancy-based technique not only using *CPR*, but also with several other standard redundancy-based measures. Furthermore, `PM-2` is more effective at handling the case where topics are not equally popular. Our results show that promoting proportionality will result in minimal redundancy, or equivalently maximal novelty, as desired by the current diversity standards.

CHAPTER 5

INFERRING QUERY TOPICS FROM REFORMULATIONS USING CLUSTERING

5.1 Introduction

As introduced in the previous chapters, the explicit approach to search result diversification (Agrawal et al., 2009; Carterette & Chandar, 2009; Santos et al., 2010a) often assumes that the set of topics associated with the query is available. Generating these topics automatically, on the other hand, has been less successful.

Notable work in this area includes that by Radlinski et al. (2010). To provide topics for a query, they propose to cluster related queries, or reformulations, from a large proprietary query log and use each group of queries to represent a topic. They show that their method can provide reasonable clusters of queries. As a result, it was used during topic development for TREC Web tracks (Clarke, Craswell, & Soboroff, 2009; Clarke et al., 2010; Clarke, Craswell, Soboroff, & Voorhees, 2011; Clarke & Craswell, 2012), in which human judges examined the output for each query and manually determined the set of probable topics as well as provided a descriptions for each of these topics. Regardless, the effectiveness of this method for automatically inferring query topics for diversification has yet to be confirmed.

Motivated by this work, we propose to cluster the reformulations for each user query generated from publicly available resources, including anchor text and Microsoft Web N-Gram Services [1]. Firstly, we show that many of the reformulations we generate for TREC queries correspond very well with the topics that the human

[1] http://research.microsoft.com/en-us/collaboration/focus/cs/web-ngram.aspx

judges identified. We then show that our approach can provide consistent topical clusters of reformulations, and diversification based on these clusters outperforms the standard relevance-based retrieval model.

Furthermore, He et al. (2012) show that the right combination of topics generated using our method and those generated from the retrieved documents and query logs can improve diversification effectiveness significantly.

The remainder of this chapter is organized as follows. Section 5.2 explains how we generate reformulations for a query. Section 5.3 describes the clustering algorithms and similarity measures that we study. Finally, Section 5.4 presents our evaluation results.

5.2 Generating Reformulations

Even though many techniques have been proposed for query reformulation, most of them aim to generate queries that are more effective for retrieval without changing the original user intent (Jones et al., 2006; Dang & Croft, 2010). Their effectiveness for providing reformulations that cover different query topics is thus unclear. Instead of using these proposed techniques, we use a rather simple method for generating reformulations from two publicly available resources: anchor text and web ngrams.

5.2.1 Anchor Text

Anchor text is known to be an effective feature for web search (Metzler et al., 2009). Previous researchers have observed the similarity between anchor text and queries (Eiron & McCurley, 2003; Dang & Croft, 2010). Therefore, we treat each anchor text as a reformulation that can potentially represent a query topic.

The web collection from which we extract the anchor text is the English portion of the ClueWeb-09 category A[2]. It contains 500 million pages in English that were

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

crawled from the web during early 2009. We extracted all pairs of anchor text and associated urls from the web pages in this collection.

Web pages are connected to one another via links, each of which is associated with some anchor text. A link is called internal if two connected pages are from the same domain and external if they come from different domains. Since most of the internal links are for navigation purposes, their associated anchor text is not very helpful. Typical examples of such anchor text are "home" and "index". As a result, we only consider external links.

In order to reduce noise, we discarded anchors that contain non-English words and those that contain navigation-triggered words such as "click", "download" and "subscribe". We also removed anchors that contain only numbers and stop words. Among the resulting anchors, we keep only those with frequency greater than 1 and that are connected to at least *two* urls. The resulting anchor text collection contains $8,215,751$ unique anchors.

For any given query, we use the top M most frequent anchor texts that contain all of its terms as its reformulations.

5.2.2 Microsoft Web N-gram Services

The Microsoft Web N-gram Services provide smoothed n-gram models built from document body, document title, anchor text and queries in the Bing query log. Each model gives the probability $P(u|n)$ of seeing an unigram u coming after an n-gram n. For each query q, we obtain the top M unigrams u with largest $P(u|q)$. Each reformulation is formed by adding u to q.

We put all reformulations generated from the two sources above into a list L. Since we aim to use each reformulation as a query topic, we keep only those with reasonably high frequency. Ideally, we can obtain this frequency from query logs. Because we rely only on publicly available resources, we approximate this frequency

by the number of times all of the query terms co-occur within a window of size 10 in a web collection. This is done with the unordered window $\#uw10(...)$ operator implemented in Indri. Finally, we order the reformulations in L by their frequency and keep only the top M, which will be the candidates for clustering.

5.3 Clustering

The list of reformulated queries generated above are then clustered into groups, each of which is considered a coarse representation of a query topic. The clustering is based on a measure of query similarity.

5.3.1 Similarity Measures

We use two types of query similarity measures in this study. The first one is based on relevance models (Lavrenko & Croft, 2001) and the second one is based on co-occurrence in passages.

5.3.1.1 Relevance Models

Since the queries are short, computing their similarity based only on the query words is not likely to be effective. Therefore, we expand each query with terms from the documents that are potentially relevant to it. Specifically, we represent a query by the relevance model (Lavrenko & Croft, 2001) estimated from the top K documents returned by the Query Likelihood retrieval model (Ponte & Croft, 1998) for this query. We choose Query Likelihood for simplicity, but other retrieval models could also be used.

Formally, let R be the set of documents retrieved for the query q and $P_{QL}(d|q)$ indicate the query likelihood score for a document $d \in R$. Let W denote the set of non-stopword terms extracted from all documents in R. A relevance model is a

distribution over all words in W. The probability that each term $w \in W$ comes from this model is given by:

$$P_q(w) = \sum_{d \in R} P(w|d) P_{QL}(d|q)$$

where $P(w|d)$ is an estimate of the probability that w can be generated from d. The idea is that a term that occurs frequently in highly ranked, and therefore probably relevant, documents is better at describing the query intent.

The similarity of two reformulations r_1 and r_2 is based on the Kullback–Leibler (KL) divergence between their relevance models $P_{r_1}(w)$ and $P_{r_2}(w)$. The KL divergence of Q from P is given by:

$$D_{KL}(P||Q) = \sum_w P(w) \log \frac{P(w)}{Q(w)}$$

To enforce symmetry, we calculate the similarity between these two reformulations as follows:

$$sim_{KL}(r_1, r_2) = \frac{1}{2} \left(D_{KL} \left(P_{r_1}(w) || P_{r_2}(w) \right) + D_{KL} \left(P_{r_2}(w) || P_{r_1}(w) \right) \right)$$

In addition, we also experiment with the cosine similarity measure as an alternative to KL divergence:

$$sim_{cos}(r_1, r_2) = \frac{\sum_w P_{r_1}(w) P_{r_2}(w)}{\sqrt{\sum_w P_{r_1}^2(w)} \sqrt{\sum_w P_{r_2}^2(w)}}$$

5.3.1.2 Co-occurrence At Passage Level

Since estimating relevance models for every reformulation is computationally expensive, we also examine a more efficient method based on passage analysis. The idea is that two queries are more similar if they co-occur often in the same text passages. Therefore, for every pair of reformulations r_i and r_j, we compute N_i and N_j – the

number of passages in which each of them occurs, and N – the number of passages in which they co-occur. The similarity between r_i and r_j is given by the Jaccard score:

$$sim(r_i, r_j) = \frac{N}{N_i + N_j - N}$$

For efficiency reasons, we approximate N with the number of times that all terms in both reformulation r_i and r_j co-occur within a window of size 20 in our document index. If two reformulations have some terms in common, these shared terms only need to occur once in this window. N_i is also approximated in a similar fashion.

5.3.2 Clustering Algorithms

We experiment with two clustering algorithms: K-Means and agglomerative clustering.

5.3.2.1 K-Means Clustering

The algorithm initializes each of the K clusters with a random reformulation. It then iteratively partitions all reformulations into K clusters in which each reformulation is assigned to the cluster that is most similar to it. The similarity between a query r_i and a cluster C_k is the average similarity between this query and all queries in the cluster:

$$sim(r_i, C_k) = \frac{\sum_{r_j \in C_k} sim(r_i, r_j)}{|C_k|}$$

where $sim(r_i, r_j)$ is the similarity measure between two reformulations as described in Section 5.3.1. The algorithm terminates when the cluster assignment for reformulations does not change.

5.3.2.2 Agglomerative Clustering

Agglomerative clustering has an advantage over K-Means in that we do not have to specify the number of clusters beforehand. The standard algorithm treats each

reformulation as a singleton cluster. It successively merges pairs of clusters that are most similar to each other until some criteria are achieved. In our experiments, the algorithm stops when the intra-cluster similarity (the average pair-wise similarity) drops below a certain threshold τ. We use complete-link to compute the similarity between two clusters C_l and C_k, which is the minimum pair-wise similarity between the queries in these clusters:

$$sim(C_l, C_k) = \min_{r_i \in C_l, r_j \in C_k} sim(r_i, r_j)$$

5.4 Experiments

In our experiments, we use queries from TREC Web Track 2009 and 2010. This query set contains 100 queries. Each query comes with an associated set of ground-truth topics identified by TREC assessors. For each query, we generate reformulations using the procedure described in Section 5.2. We evaluate the quality of these reformulations by judging how many of them correspond to the ground-truth topics. Then, we examine the topical consistency in the clusters provided by different combinations of clustering algorithms and similarity measures. Finally, we use the set of clusters generated for each query as its topic set for diversification and study its effectiveness.

5.4.1 Data Preparation and Parameter Settings

We used ClueWeb-09 category B for estimating both the frequency of reformulations and the co-occurrence statistics. The frequency of a reformulated query is the number of times all of its terms co-occur within a windows of size 10. For passage analysis, two reformulations are considered co-occurring if all of their terms co-occur within a window of size 20. Regarding τ (the intra-cluster similarity threshold) and K (the number of clusters for K-Means), we examined the clusters generated for a few queries and choose the value that provides the best results: $\tau = 0.45$ and $K = 10$.

Last but not least, we set $M = 100$ (the number of reformulations) in all of our experiments.

5.4.2 Quality of Reformulations

As mentioned earlier, we put all reformulations generated from different sources together for each query and keep only the top 100 most frequent ones. Among these reformulations, 15% are exclusively from the anchor text, 76% are exclusively from the Web N-gram service and 9% are from both sources.

In this experiment, two graduate students independently judged each of those 100 reformulations to see if it corresponds to any actual topics of the query. A reformulation is then labeled by the corresponding topic, or "none" if it does not match with any of the topics. The agreement between our two judges is 94%.

A ground-truth topic of the query is considered covered if at least one of the reformulations corresponds to it. Fig. 5.1 shows the percentage of topics (averaged across all queries) covered by the top N of the 100 reformulations with N varying from 10 to 100. In general, the reformulations covers on average about 60% of the actual topics, which is promising considering these reformulations are acquired in a very simple way. This suggests that publicly available resources are very useful for identifying topics of queries.

It is worth noting that the reformulations that do not correspond to any of the true topics are not necessarily incorrect. In fact, many of them represent valid intents that were not identified by TREC assessors. We leave the evaluation of these reformulations for future work.

5.4.3 Quality of Clusters

We now evaluate different combinations of clustering algorithms and similarity measures used to group the reformulations we have generated. We expect the techniques to be able to put reformulations with the same label into the same cluster.

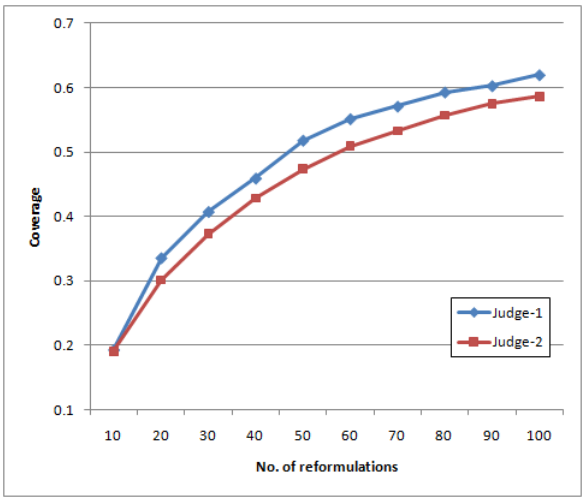


Figure 5.1: Quality of the generated reformulations in terms of how many of the actual topics of the queries they cover.

Table 5.1: Quality of the automatically generated reformulations.

		RM+Cos.	RM+KL	PS+JAC
Judge-1	Agglo.	0.64	0.67	**0.76**
	K-Means	0.5	0.55	0.59
Judge-2	Agglo.	0.63	0.7	**0.73**
	K-Means	0.48	0.55	0.57

To evaluate the quality of the generated clusters, we use the Rand index (RI), a well-known cluster quality measure. It computes the percentage of decisions that are correct and is calculated as follows:

$$RI = \frac{TP + TN}{TP + FP + FN + TN}$$

where TP (true positive) is the number of pairs of reformulations with the same labels that are put into the same cluster, TN (true negative) is the number of pairs with different labels that are put into the same cluster, FN (false negative) is the number of pairs with the same labels that are put into different clusters, and FP (false positive) is the number of pairs with different labels that are put into the same clusters. Reformulations with the label "none" are ignored in this computation

since "none" is not a topic. Table 5.1 shows the RI score that different combinations achieve.

The first thing we observe from Table 5.1 is that agglomerative clustering consistently outperforms K-Means. The reason seems to be that K-Means forces every reformulations to be in some cluster. This can result in unrelated reformulations being put into the same cluster. Once clusters contain many unrelated reformulations, the centroids of those clusters are not very different from each other, making the cluster assignment in the next iteration unreliable. Agglomerative clustering only merges two clusters if they are very similar to each other, and has a lower chance of putting reformulations into unrelated clusters.

Secondly, Table 5.1 shows that the similarity measure based on co-occurrence is consistently better than those based on relevance models. It should be noted that most of the reformulations, especially those generated from the Microsoft N-Gram Services, are different to each other by only one word. The longer the original query, the less impact the augmented word has on the relevance model. As a result, the relevance models for these reformulations are more similar than they should be. The similarity measure based on co-occurence, on the other hand, is not affected as much by the length of the original query. Two reformulations are similar as long as their augmented words co-occur with each other and with the original query. This gives the co-occurrence-based measure an advantage. Tables 5.2 presents an example of clusters generated by agglomerative clustering with the co-occurrence similarity measure for the query "satellite".

5.4.4 Diversification Effectiveness

We now investigate the effectiveness for diversification of the clusters generated using agglomerative clustering with the co-occurrence-based similarity measure. We treat the clusters generated for each of the 100 queries as the potential topics. We

Table 5.2: Example of clusters generated by agglomerative clustering for the query "satellite"

{satellite tv; satellite tv vs cable; satellite network}
{satellite radio; sirius satellite radio; xm satellite radio}
{satellite image; google maps satellite}
{satellite internet}
{weather satellite; satellite climate}
{satellite technology; satellite development}
{satellite broadband}

then discard all clusters with only one reformulation, which we assume to be very infrequent topics and thus should be ignored. This leaves some queries with only one remaining cluster, which is not interesting for diversification. Consequently, we only consider a subset of 77 queries for which our techniques can provide at least two clusters.

The retrieval collection, experimental setup and evaluation measures in these experiments are very similar those used in our previous experiments in which we evaluate our proportionality models (see Chapter 4 for details). In brief, for each of the 77 queries, we use Query Likelihood (Ponte & Croft, 1998) to retrieve an initial ranking of documents from the ClueWeb-09 category B document index. The set of associated clusters is provided as input to a diversification system. This system re-orders the input ranking to make it more diverse with respect to these topic clusters. We evaluate the final rankings by how well they cover the ground-truth topics using a variety of diversity measures. For the diversification techniques, we used `xQuAD` (Santos et al., 2010a), `PM-1` and `PM-2`. All model parameters are selected via 2−fold cross validation.

These diversification techniques assume that the estimate of how relevant a document is to a topic is available. Each of our topics is a cluster of queries. One can certainly estimate the relevance of a document to a cluster using the average of its relevance to each of the query in this cluster. However, this requires running mul-

Table 5.3: The effectiveness of our topics for diversification. No statistical significance is observed with respect to the baseline Query Likelihood.

	CPR	α-NDCG	ERR-IA	Prec-IA	S-Recall	NRBP
Query-likelihood	0.3669	0.2637	0.1644	0.1113	0.4107	0.1332
xQuAD	0.3598	0.2601	0.1620	0.1169	0.4052	0.1299
PM-1	**0.3943**	**0.2944**	**0.1961**	**0.1306**	**0.4189**	**0.1685**
PM-2	0.3703	0.2828	0.1888	0.1157	0.4010	0.1641

tiple queries over the index. For simplicity and efficiency, we estimate it differently as follows. We concatenate all queries in each cluster to form a "document", from which we construct a unigram language model. Finally, we use Indri's weighted query representation of this model as the topic description. The relevance of a document to a cluster is its query likelihood score with respect to the corresponding weighted query. We consider all topics to be equally popular.

The results are presented in Table 5.3. Although the performance of `xQuAD` is lower than the baseline, both `PM-1` and `PM-2` manage to provide improvement with nearly all measures. While no statistically significant differences are observed between these techniques and the baseline, the results still suggest that the topics generated with our technique can be beneficial.

Interestingly, `PM-1` is the best performing approach despite its naive assumption that a document only belongs to one topic. We believe the reason is as follows. Recall that only a portion of the topics we generated are in the ground-truth set. We will refer to those that are not in this set as the misaligned topics. Due to our evaluation setup, a system is penalized if it promotes documents for these misaligned topics, even if they represent valid user intents. `PM-2` and `xQuAD` generally favor documents that are relevant to multiple topics. It is possible that they select a document that is relevant to multiple misaligned topics over those that are relevant to a single ground-truth aspect. `PM-1`, on the other hand, scores a document based solely on how well it covers the most under-represented topic. At the iteration where a ground-truth topic

is the most under-represented, PM-1 always selects the best document for this topic regardless of its relevance to others. Thus, PM-1 is less affected by the misalignment problem.

5.5 Summary

In this chapter, we have shown that reformulations for queries obtained from publicly available resources such as anchor text and Microsoft Web N-Gram Services can provide coverage for a broad range of query topics. We tested different combinations of clustering algorithms and query similarity measures for grouping the reformulations that are topically related. We found that agglomerative clustering consistently outperforms K-Means and the similarity measure based on co-occurence is not only more efficient but also more effective than the similarity that is based on relevance models. Additionally, we demonstrated that the sets of clusters generated by this combination are topically consistent and effective for diversification.

CHAPTER 6

TERM LEVEL SEARCH RESULT DIVERSIFICATION

6.1 Introduction

We have described previous research on techniques for generating query topics for diversification. Carterette and Chandar (2009) applied Latent Dirichlet Allocation (Blei et al., 2003) on a set of documents retrieved for a query to learn its topics. The same authors also used the k-nearest neighbor algorithm to cluster the same set of documents and use the relevance model (Lavrenko & Croft, 2001) estimated from each resulting cluster to represent a topic. Instead of using retrieved documents as the source for topic extraction, Radlinski et al. (2010) proposed using query logs. Their method clusters related queries in a large proprietary log and uses each cluster to represent a query topic. We described a similar query clustering technique in Chapter 5 that uses anchor text and web n-grams instead of query logs. Dou et al. (2011) combined multiple sources of information to form topics including clusters of documents, anchor text and query logs.

The success of these techniques, however, has been quite limited. Carterette and Chandar (2009) evaluate their methods on a small newswire collection. It is unclear if these results can be generalized to noisy web collections. The technique by Radlinski et al. (2010) is demonstrated to provide topically consistent clusters of queries, but these clusters have not been evaluated for diversification effectiveness. Although Dou et al. (2011) show that their generated topics are beneficial to the diversification of web documents, their evaluation is done using a small number of queries. Furthermore, they report the results on the same data that was used to tune the model parameters.

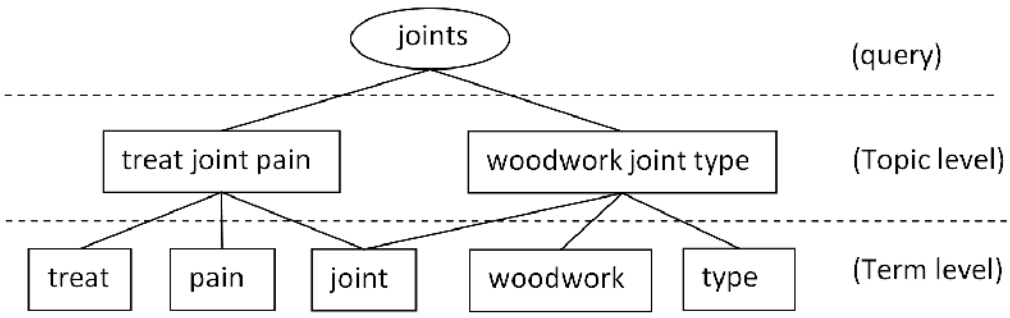


Figure 6.1: Two different levels for diversification: topic level and term level.

It is thus unclear how well their optimal values can generalize. To the best of our knowledge, only the most recent methods show consistent improvement over non-diversification retrieval baselines on web corpora. This includes the techniques by He et al. (2012) and Santos et al. (2013). He et al. (2012)'s approach infers query topics from multiple sources of information using a regularized topic modeling approach (Cai et al., 2008), while Santos et al. (2013) use learning to rank to select related queries from a query log.

Human descriptions or labels for topics usually take the form of a coherent group of terms. Fig. 6.1 shows an example TREC query, *joints*, with the description provided by TREC assessors for two of its topics: *treat join pain* and *woodwork joint type*[1]. Our experiments in Chapter 4 have shown that using these human-created descriptions, PM-2 can effectively diversify the result rankings.

The existing work in topic generation can be seen as attempts to generate substitutes for these concise descriptions. A distribution of terms (Carterette & Chandar, 2009), as a matter of fact, is a very coarse substitute. Although anchor text and queries are more succinct and thus are potentially better topic representation, identifying those that represent a particular topic (Radlinski et al., 2010; Dou et al., 2011; Santos et al., 2013) can be challenging, not to mention they might not exist for all

[1] We only show the keywords for brevity. The full descriptions for these two topics are *joint pain and how to treat it* and *different types of joints used in woodworking.*

topics. Based on the limited success with these methods, we argue that generating for each query a set of concise topic descriptions, each of which is a coherent group of terms, is very difficult.

Instead of creating another technique that attempts to generate such descriptions, we question the necessity of this representation for improving diversity in search results. Our hypothesis is that being able to identify the important terms that make up these topics will be sufficient to achieve improvement in diversity. Our intuition is based on the nature of existing diversification techniques such as `PM-2` and `xQuAD` that favor documents that are relevant to multiple topics. As an example, we could provide `PM-2` with a set of terms including *treat*, *joint*, *pain*, *woodwork*, and *type* as though each term was a topic. A document about *treat join pain* should contain many occurrences of these words and would have a high "relevance" for each term. Similarly, a document about *woodwork joint type* should appear relevant to *woodwork*, *joint* and *type*. Given this, `PM-2` should select these two documents over those that provide general information about a particular type of joint in human body, which would only appear relevant to the term *joint*. In other words, we believe that using this set of five terms as the "topics" would be sufficient for existing techniques to promote diversity with respect to the two higher-level topics.

Consequently, we experiment with this simpler term-based representation for a set of query topics. Instead of modeling each topic as a group of terms, we use these terms directly without their grouping structure (Section 6.2). This means that for the query *joint*, we will model all of its topics using a set of five terms: *treat*, *joint*, *pain*, *woodwork* and *type*, as shown in Fig. 6.1. This set of terms is provided as input to a diversification system which treats each of them as a topic. We refer to this approach as *term level search result diversification*.

In the next chapter, we demonstrate empirically that when we know the ground-truth set of topics associated with the query, diversifying a result ranking using just

the terms from these topics can significantly improve topical diversity. In fact, our approach is quite competitive to using the set of topics itself. This shows that while the grouping structure, assuming that it can be accurately identified, can provide some additional benefit to diversification compared to just having the terms, the terms are sufficient to improve diversity in the final ranking. This simplifies the task of finding a set of topics, which has proven difficult, into finding an appropriate set of terms. Therefore, instead of trying to generate a precise description for each of the query topics, we only need to identify a set of terms that provide good coverage for these topics. This is, in fact, the main task for term-based multi-document summarization (Sanderson & Croft, 1999; Lawrie et al., 2001; Lawrie & Croft, 2003).

Consequently, we propose to use `DSPApprox`, a greedy algorithm from the document summarization literature for identifying terms for diversification from the initial ranking of documents (Lawrie & Croft, 2003) (Section 6.3). Our results show that this relatively simple method significantly outperforms many existing approaches for estimating the full topic structure from the same data on a wide ranges of both relevance and diversity measures. To the best of our knowledge, at the time our approach was introduced, it was the first to provide statistically significant improvement over standard relevance-based retrieval models in terms of both relevance and diversity measures, without relying on any external data source or manually created topic set.

Interestingly, finding a set of terms is, in fact, a diversification problem by itself. We show the connection between the term generation problem and the document diversification problem as well as analyze the similarity and differences betweeen `DSPApprox` and existing document diversification techniques (Section 6.3). This enables the application of those existing methods to the task of generating a diverse set of terms that are beneficial to document diversification.

The last part of this chapter explores several sources of information from which we can extract highly descriptive terms (Section 6.3.2).

6.2 Term Level Search Result Diversification

In this section, we first summarize the problem of diversification at the topic level, which has been formally described in Chapter 3. We then introduce our term level approach and provide some intuition for why one can expect it to promote topical diversity in search results.

6.2.1 Topic Level Diversification

Let q indicate a user query and $T = \{t_1, t_2, ..., t_n\}$ indicate the set of topics for q. Let $W = \{w_1, w_2, ..., w_n\}$ denote the weights for each of the topics $t \in T$. These weights can be interpreted as the importance or popularity of each topic. In addition, let $R = \{d_1, d_2, ..., d_m\}$ indicate a ranked list of documents initially retrieved for q and $P(d|t)$ denote some probabilistic estimate of d's relevance to a topic t. The task of topic level diversification is to select a subset of R using $\{T, W, P(d|t)\}$ to form a diverse ranked list S of size k.

It is worth noting that the representation of the topics $T = \{t_1, t_2, ..., t_n\}$ will determine the relevance measure $P(d|t)$. For example, if T is a set of short textual descriptions (e.g. queries), $P(d|t)$ is often the relevance score of d to t given by some retrieval models. In this work, we assume that each true topic of the query can be represented as a set of terms, which we will refer to as *topic terms*. From this moment on, we will use the word *"topic"* to refer to this particular representation, and *"topic terms"* (or *"terms"* for short when there is no confusion) to refer to the words that make up a topic.

6.2.2 Term Level Diversification

We will reuse the example query in Fig. 6.1 to explain the idea of term level diversification. We assume that the query *joint* has two topics: *treat join pain* and *woodwork joint type*. Our experiments in Chapter 4 have shown that if we can correctly represent these query topics, `PM-2` can effectively diversify the result rankings.

Generating topic descriptions that are as concise as *treat join pain* and *woodwork joint type*, however, is extremely difficult. Identifying the individual topic terms – *treat*, *joint*, *pain*, *woodwork*, and *type*– on the other hand, is a relatively well studied problem in the area of multi-document summarization.

Assuming we can identify these terms, our term level approach will use them with diversification techniques such as `PM-2`, which essentially treats each of these terms as a topic and perform diversification with respect to these "topics". Using our example, this means that we give the five terms – *treat*, *joint*, *pain*, *woodwork*, and *type*– to `PM-2`. A natural question to ask is, what would the result ranking looks like? As far as topical diversification is concerned, how can one expect the search results to be diverse with respect to the two topics when `PM-2` only uses a set of terms? In the remainder of this section, we first present a formal statement of this problem as well as our assumptions, and then answer these questions.

6.2.2.1 Problem Statement

Diversification at the term level is very similar to the topic level. Let $t_i = \{t_i^1, t_i^2, ..., t_i^{|t_i|}\}$ be the set of terms for topic t_i. Instead of diversifying R using the set of topics $T = \{t_1, t_2, ..., t_n\}$, we propose to perform diversification using $T' = \{t_1^1, t_1^2, ..., t_1^{|t_1|}, ..., t_n^1, t_n^2, ..., t_n^{|t_n|}\}$, in effect treating each t_i^j as a topic. Consequently, any technique that has been proposed for topic level diversification can be use to perform term level diversification.

6.2.2.2 Assumptions

Our approach makes a few assumptions. First, any given set of topic terms is assumed to have an underlying set of latent topics. For example, the set of *treat*, *joint*, *pain*, *woodwork*, and *type* has at least two latent topics: *treat join pain* and *woodwork joint type*.

Second, if a document is highly relevant to a topic, it is highly relevant to all of the corresponding topic terms (which means it will contain many instances of those words). For instance, a document is considered highly relevant to *treat*, *joint* and *pain* if it is relevant to the *treat join pain* topic.

Last but not least, if a document is more relevant to a topic t_i than it is to t_j, it is more relevant to the terms that make up t_i than those representing t_j. This means that if a document is more relevant to *treat join pain* than it is to *woodwork joint type*, it is also more relevant to *treat* and *pain* than it is to *woodwork* and *type*. We ignore the common term (*joint*) between topics for the ease of explanation.

6.2.2.3 How It Works

Consider an initial ranking R with three documents. While d_1 and d_2 are relevant to *treat join pain* and *woodwork joint type* respectively, d_3 provides general information about a particular type of joint in human body (e.g. elbow joint), thus is *not* relevant to the user' information needs. Under our assumptions, d_3 is highly relevant only to *joint*. While both d_1 and d_2 are also relevant to *joint*, d_1 is substantially more relevant to *treat* and *pain* than it is to *woodwork* and *type*, and d_2 will be more relevant to the latter two. This is visualized in Figure 6.2.

We will now explain how `PM-2` works with this set of five topic terms as "topics". Recall that at each iteration, `PM-2` computes a quotient for each term. Assuming the terms have equal weight, it is easy to verify that in the first iteration, the quotient for all five terms is 0.2. Setting $\lambda = 0.5$ for simplicity, each of the three candidate documents will be scored by:

$$score(d) = \sum_{t \in T} q_t P(d|t) \tag{6.1}$$

where T is now the set of topic terms. The summation indicates the key property of `PM-2` (and many other diversification techniques): it favors documents that are

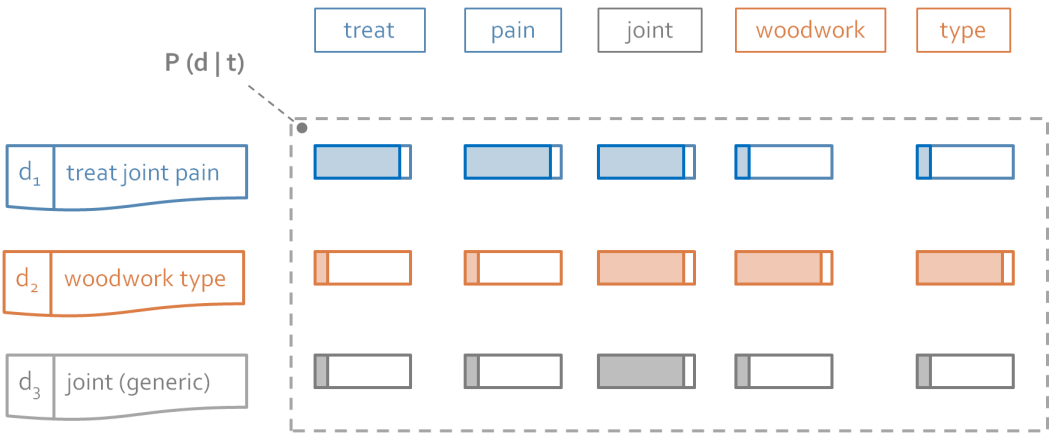


Figure 6.2: The relevance between each candidate documents to each of the topic terms being used for diversification.

relevant to multiple "topics". Since d_1 and d_2 are relevant to three terms while d_3 is only relevant to *joint*, d_1 and d_2 will receive a higher score. Regardless which document (d_1 or d_2) is selected, the other will be chosen in the next iteration over d_3 for the same reason. Interestingly, although `PM-2` only takes as input a set of topic terms, it manages to return relevant documents for both underlying topics due to its tendency to promote multi-topic documents. This suggests that diversifying a result ranking using a set of terms can improve diversity with respect to the topics underlying these terms.

Notice in the example above the presence of a third latent topic that is not relevant to the user's intents: general information about elbow joints. `PM-2` is able to avoid choosing d_3 because this topic is represented only by one term in the set given to `PM-2` (*joint*) while each of the two query topics consists of three terms. This indicates the necessary condition for a query topic to be covered in the results: it must have sufficient presence in the set of topic terms provided to the diversification system.

It is worth noting that this description provides an intuition of what the result ranking generated by our approach looks like, and how it accounts for diversity with respect to the query topics given that it only takes as input a set of terms. For

the ease of explanation, we used a rather restricted setting where each document is assumed to belong to one topic and λ is set to 0.5 for `PM-2`. In practice, documents are multi-topics and model parameters are tuned in some way. Although we believe our reasoning should generalize at least to some extent, it is extremely difficult to support this claim analytically. Thus, we will rely on empirical evaluation to verify the validity of our approach.

6.2.2.4 Choice of $P(d|t)$

As mentioned earlier, diversification techniques assume $\{T, W, P(d|t)\}$ as inputs and the choice of T will determine $P(d|t)$. An obvious choice for $P(d|t)$ for term level diversification is $P(t_i^k|d)$, the probability that the document d generates the topic term t_i^k. This is, however, highly problematic. At the term level, in addition to those "true" query topics that are now latent, there are also "false" latent topics formed by the wrong combinations of terms. Using the five example terms *treat*, *joint*, *pain*, *woodwork* and *type*, *pain caused by woodworking* can be one of these "false" topics. In the context where we identify topic terms for a query automatically, some of them might be generic and ineffective. As the number of bad terms increases, the number of "false" topics will grow exponentially. Combined with the fact that there are likely to be many non-relevant documents in the baseline ranking, term diversification under the effects of these "false" topics could end up promoting non-relevant documents.

Assuming any document that is relevant to a true query topic should be relevant to the query itself, we propose to incorporate the relevance of a document to the query into the estimation of $P(d|t)$. Let $\{q_1, q_2, ..., q_n\}$ be the set of terms of the query q. $P(d|t_i^k)$ is estimated as follows:

$$P(d|t_i^k) \simeq (P(t_i^k|d)P(q|d))^{\frac{1}{|t_i^k|+|q|}} = (P(t_i^k|d)\prod_{q_j \in q} P(q_j|d))^{\frac{1}{|t_i^k|+|q|}}$$

which is essentially the Query Likelihood score (Ponte & Croft, 1998) for ranking d with respect to the query $\{t_i^k, q_1, q_2, ..., q_n\}$ normalized by the query length to avoid biased towards terms with fewer words (i.e. terms can include both unigrams or phrases). In the case where all terms have the same length, the normalization is certainly not necessary.

6.3 Automatic Extraction of Topic Terms

In this section, we present how `DSPApprox`, an algorithm proposed by Lawrie and Croft (2003) for hierarchical multi-document summarization, can be used to extract a diverse set of topic terms for our diversification approach. We then explain why the task of term generation can be considered a diversification problem. Next, we analyze the similarity and differences between `DSPApprox` and existing document diversification techniques, which enables the possibility of using these existing techniques for effective term generation.

6.3.1 `DSPApprox`

The goal of `DSPApprox` is to select from a collection of documents a small set of highly representative terms that best summarize the topics covered by the documents. This algorithm is applied recursively, resulting in an hierarchical structure of topic terms.

Since the documents in an initial ranking retrieved for a query have high probability of being relevant to some of this query's topics, we consider these documents a valuable source for extracting terms that can represent those query topics. As a result, we apply `DSPApprox` on these documents. There are other potential sources for topic term extraction as well, which we will describe in the next section. Since we only need a single diverse set of topic terms, we only use this algorithm to generate a single level instead of a hierarchy of terms.

The algorithm first identifies a set of *vocabulary* words from the collection of documents, from which it forms a set of more specific *topic terms*. It then measures for these terms their *topicality* and how well they predict the occurrences of other terms. Finally, it greedily selects a subset of topic terms, aiming to maximize both their topicality and their coverage of the vocabulary.

Vocabulary Identification. We consider as vocabulary all terms that (1) appear in at least two documents, (2) have at least two characters and (3) are not numbers. In our experiments, we test two types of terms separately: unigrams and phrases. We use a very simple method for phrase extraction. We scan through the words in each document from the beginning to the end. For each word w_i, we select the longest consecutive sequence $p = \{w_i, w_{i+1}, ..., w_k\}$ such that p matches the title of a Wikipedia page and $p \cup \{w_{k+1}\}$ does not. p is then selected as a vocabulary phrase and the process repeats at w_{k+1}.

Topic Term Identification. All vocabulary terms that co-occur with any of the query terms within a proximity window of size w are selected as topic terms.

Topicality and Predictiveness. The topicality of a term measures how informative it is at describing the set of documents. To compute topicality, a relevance model $P_R(t|q)$ (Lavrenko & Croft, 2001) is first estimated from the initial set of documents R:

$$P_R(t|q) = \sum_{d_i \in R} P(t|d_i)P(d_i|q)$$

where $P(t|d)$ is the probability that d_i generates the term t and $P(d_i|q)$ is relevance of d_i to the query. The topicality $TP(t)$ of a term t is estimated as its contribution to the Kullback–Leibler divergence between this relevance model and the language model for the entire retrieval collection $P_c(t)$:

$$TP(t) = P_R(t|q) log_2 \frac{P_R(t|q)}{P_c(t)}$$

Table 6.1: Example output of `DSPApprox` for the query "joints" (topic number 82). Some of the original TREC subtopics for this query are also provided for comparison.

TREC Sub-topic	`DSPApprox`[Unigram]	`DSPApprox`[Phrase]
1) joints in human body	spine	elbow joint
	articulate	knee joint
2) woodworking joints types	miter	miter joint
	planter	miter box
3) treat joint pain	symptom	joint pain
	grease	joint anti inflammatory

It is equivalently t's contribution to the clarity score of the query q (Cronen-Townsend et al., 2002).

Predictiveness, on the other hand, measures how much the occurrence of a term predicts the occurrences of others. Let $P_w(t|v)$ indicate the probability that a term t occurs within a window of size w of another term v and V_t indicate the set all such v with respect to t. The predictiveness of t is estimated as follows:

$$PR(t) = \frac{1}{Z} \sum_{v \in V_t} P_w(t|v) \tag{6.2}$$

where Z is some normalization factor. We set it to the size of the vocabulary.

Greedy Algorithm. Pseudo-code for this algorithm is presented as Algorithm 7. It iteratively selects terms from the candidate topic term set T. The utility of each term is the product of its topicality and predictiveness. At each step, the algorithm selects the topic term $t^* \in T$ with maximum utility. Then, it decreases the predictiveness of other topic terms that predict the same vocabulary. This makes sure topic terms that cover the uncovered part of the vocabulary will emerge for selection in the next iteration. The algorithm stops once the utility of all candidate topic terms reaches 0, indicating that all vocabulary has been covered. Some example topic terms (both unigrams and phrases) generated by `DSPApprox` for the query *joints* are shown in Table 6.1.

Algorithm 7 `DSPApprox` for identifying topic terms.

$V = \{v_1, v_2, ..., v_n\}$: the set of vocabulary
$T = \{t_1, t_2, ..., t_m\}$: the set of candidate topic terms
V_t: set of vocabulary words occurring within a window to t
$P_w(t|v)$: co-occurrence (within window of size w) statistics
Compute topicality $TP(t), \forall t \in T$
Compute predictiveness $PR(t), \forall t \in T$
S : **the output diverse set of topic terms**
$PREDV$: vocabulary that has been predicted by S
$S \leftarrow \emptyset$
$PREDV \leftarrow \emptyset$
while $PREDV \subset V$ **and** $|T| > 0$ **do**
    $t^* \leftarrow \arg\max_{t \in T} TP(t) \times PR(t)$
    $S \leftarrow S \cup t^*$
    $T \leftarrow T \setminus \{t^*\}$
    for all $v \in V_{t^*} \cap PREDV$ **do**
        for all $t \in T$ **do**
            $PR(t) \leftarrow PR(t) - P_w(t|v)$
        end for
    end for
    $PREDV = PREDV \cup V_{t^*}$
end while

6.3.2 Topic Term Extraction as a Diversification Problem

Interestingly, generating a set of topic terms as done by `DSPApprox` can be considered a diversification problem. Similar to diversifying a result ranking to provide better coverage for multiple query topics, `DSPApprox` diversifies a set of topic terms to provide better coverage for a set of vocabulary words. Document diversification techniques such as `PM-2` and `xQuAD` greedily selects the documents that are relevant to the query and can provide coverage for the topics that those selected previously fail to provide. `DSPApprox` iteratively selects the topic terms that are highly topical and can predict the occurrences of the part of the vocabulary that have not been covered by the terms selected earlier. This suggests that document diversification methods could be used to generate a diverse set of topic terms from some given source of information, which can then be used to perform document diversification.

So, what is the difference between `DSPApprox` and techniques such as `PM-2` and `xQuAD`? To answer this question, let us revisit how they work. Recall that `DSPApprox`

Algorithm 8 `DSPApprox` for identifying topic terms.

1: S : the output diverse set of topic terms
2: $S \leftarrow \emptyset$
3: **while** $|T| > 0$ **do**
4: $\quad t^* \leftarrow \arg\max_{t \in T} \sum_{v \in V_t} TP(t) \times P(t|v)$
5: $\quad S \leftarrow S \cup t^*$
6: $\quad T \leftarrow T \setminus \{t^*\}$
7: $\quad V_t \leftarrow V_t \setminus V_{t^*}, \forall t \in T$
8: **end while**

assumes that all vocabulary words have equal weight and computes the predictiveness of each topic term t as the sum $\sum_{v \in V_t} P(t|v)$ (Formula (6.2)). At each iteration, after a topic term t^* is selected and put into the result set S, the predictiveness of each of the remaining topic terms t is reduced by an amount equal to $\sum_{v \in V_{t^*} \cap PREDV} P_w(t|v)$ where $V_t^* \cap PREDV$ indicates the set of vocabulary words that the term t^* predicts (i.e., $P(t^*|v) > 0$) while those in S do not.

This procedure can be explained a bit differently, as shown in Algorithm 8. At each iteration, `DSPApprox` selects the best term using the following objective function:

$$t^* \leftarrow \arg\max_{t \in T} \sum_{v \in V_t} TP(t) \times P(t|v)$$

Then, after a term t^* is chosen and put into the result set S, all vocabulary words from C_{t^*} *are removed from consideration*:

$$V_t \leftarrow V_t \setminus V_{t^*}, \forall t \in T \setminus S$$

The process then repeats until termination.

Let us now revisit `PM-2` with $\lambda = 0.5$, whose objective function is as follows:

$$d^* \leftarrow \arg\max_{d \in R} \sum_{t \in T} q_t \times P(d|t)$$

Recall that PM-2 assumes that $P(d|t)$ has taken $P(d|q)$ into account. For simplicity, let us assume that this is done by multiplying these two quantities:

$$d^* \leftarrow \arg\max_{d \in R} \sum_{t \in T} q_t \times P(d|q) \times P(d|t)$$

Applying PM-2 to the topic term diversification problem by substituting the notion of topic t with vocabulary v, document d with topic term t, $P(d|q)$ for topicality $TP(t)$ and $P(d|t)$ with predictiveness $P(t|v)$, this function can be rewritten as:

$$t^* \leftarrow \arg\max_{t \in T} \sum_{v \in V_t} q_v \times TP(t) \times P(t|v)$$

Note that the quotient q_v of a vocabulary word gets smaller as the number of selected topic terms $t \in S$ that predicts v increases.

The difference between DSPApprox and PM-2 is now rather obvious: after a topic term t^* is selected, while PM-2 downweights the vocabulary words that t^* predicts by some amount, DSPApprox completely disregards these words in future iterations. One can verify easily that xQuAD and DSPApprox have the same difference. Therefore, document diversification methods can be seen as more lenient with discounting the importance of the topics that have been covered to some extent, whereas DSPApprox is substantially more aggressive.

In the context of search result diversification, one cannot estimate the relevance of a document to each topic with perfect accuracy $P(d|t)$. Ignoring a topic t as soon as a document with $P(d|t) > 0$ is selected can be disastrous. This approach will also fail to provide proportional result rankings. The aggressive nature of DSPApprox, as a result, makes it inapplicable for this task. The effectiveness of techniques such as PM-2 and xQuAD for generating a diverse set of topic terms, on the other hand, remains to be seen. As a result, we will empirically evaluate their applicability to this task.

6.4 Information Sources for Term Extraction

The `DSPApprox` technique described above was proposed for summarizing a set of documents. Naturally, we will first apply it to the ranking of documents initially retrieved for the query. Diversification techniques are then employed to diversify this same document ranking with respect to the set of topic terms `DSPApprox` provides.

Beside the initial ranked documents, there are other sources of information that can be potentially beneficial for the term extraction process. In this dissertation, we consider query logs, anchor text, Wikipedia, and Freebase as possible information sources.

6.4.1 Query Logs

User queries extracted from search logs have been proven valuable for identifying user intents (Radlinski et al., 2008). Therefore, we consider query logs as a source for finding topic terms. Specifically, we use both the AOL and the MSN query logs. The AOL log has approximately 36 million queries while the MSN log has about 15 million. These logs are used as follows. For a given query, we obtain the top K most relevant queries from each log using some standard retrieval model. These queries are treated as a "document" set on which we run `DSPApprox` to acquire the topic terms.

6.4.2 Anchor Text

In Chapter 5, we have shown how anchor text can be used to infer query topics. In addition, prior research has recognized the similarity between anchor text and user queries (Eiron & McCurley, 2003). More recently, Dang and Croft (2010) found that anchor text can be used to effectively simulate these queries for the task of query reformulation. This makes anchor text another potentially useful source for identifying topic terms. These terms are generated for each query by running `DSPApprox` on the top K retrieved anchor texts for this query.

6.4.3 Wikipedia

Wikipedia has proven to be a very reliable source for effective query expansion (Y. Xu et al., 2009). In fact, Bendersky et al. (2012) has demonstrated that, compared to the expansion terms provided by the retrieval collection, those obtained from Wikipedia usually describe different aspects of the query, thus combining them improves diversity in the result rankings. Consequently, we also apply `DSPApprox` on the top K documents retrieved from Wikipedia to extract topic terms for diversification.

6.4.4 Freebase

Freebase (Bollacker et al., 2008)[2] is a large publicly available knowledge base that contains rich structured information about real world entities and the facts associated with them. It is designed to provide coverage for highly diverse and heterogeneous data. For example, one of the TREC queries in our test set is *"defender"*. Two of the topics associated with this query, as identified by TREC assessors, are *"Windows Defender"*, the anti-spyware program, and *"Land Rover Defender"*, the sport-utility vehicle. Both *"Windows Defender"* and *"Land Rover Defender"* are entities in Freebase. This makes it a potentially valuable source for the extraction of topic terms.

Each entity in Freebase has a name (e.g. *"Windows Defender"*), a longer description and a variety of other attributes, together with its relationship to other entities. In this work, we use only the description of each entity. In particular, we retrieve the top K descriptions that are most relevant to the query for extraction.

For all four resources above, we use Query Likelihood as the retrieval model. Furthermore, we only retrieve those "documents" that contain all of the query words to ensure the quality of the extracted terms.

---

[2]http://www.freebase.com/

6.5 Summary

We have introduced a new approach to topical diversification: diversification at the term level. Existing work models a set of topics for a query, where each topic is a group of terms. While this representation is intuitive and beneficial, it makes the task of generating query topics automatically very difficult. Instead, we propose to model the topic terms directly. Our hypothesis is that being able to identify the important topic terms, which is a relatively well studied problem in the document summarization literature, is sufficient for improving diversity in a result ranking. We have also provided an intuitive justification for the fact that although our approach only takes as input a set of topic terms, it can account for the diversity with respect to the topics underlying these terms. We will empirically evaluate the validity of our approach in the next chapter.

Our term level approach to diversification effectively reduces the task of finding a set of query topics, which has proven difficult, into finding a set of topic terms. Consequently, we propose to use `DSPApprox`, a greedy algorithm from the literature of multi-document summarization (Lawrie & Croft, 2003) to identify a diverse set of terms (unigrams and phrases). Furthermore, we have shown that `DSPApprox` is indeed a diversification algorithm itself. We also presented our analyses of the similarity between `DSPApprox` and existing document diversification methods, which enables the use of these existing techniques for term generation. Last but not least, we have explored several sources of data for finding effective topic terms. In the next chapter, we will compare the effectiveness for search result diversification of the topic terms generated using these techniques with the topics generated by existing work as well as the benefits of those information sources.

CHAPTER 7

EVALUATION OF TERM LEVEL DIVERSIFICATION

In the previous chapter, we introduced our term level approach to search result diversification. Instead of inferring a set of latent topics for each query, each of which is represented by a group of terms, we extract these terms directly. This set of topic terms is provided to a diversification technique which treats each of them as a topic to determine coverage in the ranked list. We then presented how `DSPApprox`, a technique proposed for document summarization (Lawrie & Croft, 2003), can be used to generate these topic terms automatically. In addition, we showed that document diversification methods can be used for the task of term generation since it is also a diversification problem. Finally, we described alternative information sources from which we can extract effective topic terms.

In this chapter, we evaluate this approach and the associated techniques with the aim of answering the following questions:

- In a controlled environment where we have the set of concise descriptions for the ground-truth topics associated with each query, can using only the terms from these descriptions for diversification improve diversity in the final result ranking? This provides insight into the effectiveness of our term level approach.

- How does the term level approach to diversity compare with the conventional topic level approach? (i.e. using the ground-truth topics as done in Chapter 4)

- In the more realistic situation where topic terms have to be generated automatically, can the terms provided by `DSPApprox` help diversification? Furthermore,

how does this approach compare to the topics generated using existing techniques? Not only will this help us understand the effectiveness of the terms `DSPApprox` generates, it provides a practical comparison between diversification at the term level and the topic level.

- Since document diversification techniques can be used to generate a diverse set of terms, how effective are they compared to `DSPApprox`?

- Among the sources that we investigate (documents in the initial rankings, the AOL and MSN query logs, anchor text, Freebase, and Wikipedia pages), which of them are useful? Does combining terms from these sources provide any additional benefits?

- Regarding all of the questions above, do the results depend on which retrieval model is used to generate the initial ranking of documents?

We will first explain our experimental setup in Section 7.1 and then answer each of these questions with extensive analyses in the following sections.

7.1 Experimental Setup

Our setup for retrieval experiments is the same as in previous chapters. We use ClueWeb-09 category B as the retrieval collection. Our query set consists of 200 queries from Web Track 2009-2012 (Clarke, Craswell, & Soboroff, 2009; Clarke et al., 2010; Clarke, Craswell, Soboroff, & Voorhees, 2011; Clarke & Craswell, 2012). The collection is stemmed using the Krovetz stemmer (Krovetz, 1993). Stopword removal is only performed on the query using a small stopword list. A diversification system works by first using a relevance-based retrieval model to obtain an initial ranking of documents for each query. It then re-orders the top 50 documents in this ranking to provide a more diverse result list. We consider four models for the first pass retrieval: Query Likelihood (`QL`) (Ponte & Croft, 1998), Sequential Dependence

Model (SDM) (Metzler & Croft, 2005), Relevance Model (RM) (Lavrenko & Croft, 2001) and Coordinate Ascent (CA) (Metzler & Croft, 2007). Following Bendersky et al. (2010), spam filtering and the stop-word to nonstop-word ratio is incorporated as follows:

$$P(d|q) = \begin{cases} P_M(d|q) & \text{if } S(d) \geq 60 \text{ and } \sigma(d) \geq 0.1 \\ 0 & \text{otherwise} \end{cases}$$

where $P(d|q)$ is the final relevance score for document d, $P_M(d|q)$ is the retrieval score a baseline model M assigns to document d, $S(d)$ indicates the confidence that d is not a spam page (Cormack et al., 2011) and $\sigma(d)$ is the stopword-to-non-stopword ratio.

For diversification techniques, we employ xQuAD (Santos et al., 2010a), an effective redundancy-based technique, and our proportionality model PM-2. These two techniques assume that the estimate of a document relevance to a topic, $P(d|t)$, is available. For the topic level approach, we treat the description of a topic t as a query and use the query likelihood score $P_{QL}(d|t)$ as the relevance of d to this topic. For the term level approach, $P(d|t)$ is estimated as described in Chapter 6, which we restate for convenience:

$$P(d|t) \simeq \left(P\left(t|d\right) \prod_{q_j \in q} P\left(q_j|d\right) \right)^{\frac{1}{|t|+|q|}}$$

where $P(t|d)$ is the probability that the document d generates the term t, $q_j \in q$ is a term in the query. For PM-2, for both the topic level and term level approach, we combine $P(d|t)$ with the relevance of each document to the query $P(d|q)$ in a weighted manner as before. We do not do this for xQuAD, whose framework already takes $P(d|q)$ into account. Regarding model parameters, all of them are selected via 5-fold cross-validation to maximize α-*NDCG@20*, which is one of the measures we use in our evaluations. We use Fisher's randomization test for statistical significance testing.

In addition to α-*NDCG*, all systems are evaluated using a variety of other standard diversity measures, including *ERR-IA*, *NRBP*, *S-Recall* and *Precision-IA*, our proportionality measure *CPR*, and two traditional relevance measures *NDCG* and *ERR*. Each metric is computed using the top 20 retrieved documents from each ranking to be consistent with the official TREC evaluations (Clarke, Craswell, & Soboroff, 2009; Clarke et al., 2010; Clarke, Craswell, Soboroff, & Voorhees, 2011; Clarke & Craswell, 2012).

7.2 Effectiveness of Term Level Diversification

In this section, we first evaluate the effectiveness of our term level diversification approach. Each of the 200 queries is associated with a set of ground-truth topics identified by TREC assessors. Thus, each of these topic descriptions can be considered an optimal group of terms. We then discard this grouping structure and put all of the resulting unigram terms into a set for each query. We provide the topic set and the corresponding term set to the same diversification technique and evaluate the diversity in the result ranking. The diversity effectiveness of a retrieval run indicates the quality of the input set of topics and terms.

In addition, we have demonstrated earlier that related queries provided by a commercial search engine are quite effective for diversification, which is consistent with prior work (Santos et al., 2010a). These queries too can be considered good underlying topics for the original query. As a result, we also use the set of related queries and the corresponding set of terms to evaluate our term level approach. It is worth noting that the query set in this experiments only contains 190 out of the 200 queries for which the search engine provides at least two suggestions.

We first present our results and analyses in the case where Query Likelihood (`QL`) is used to retrieve the initial ranking. We will then discuss the case where we substitute

QL with the three more effective models: Sequential Dependence (SDM), Relevance Models (RM) and Coordinate Ascent (CA).

7.2.1 Query Likelihood (QL)

Table 7.1 compares our term level diversification approach to the topic level alternative using both topic sets and both diversification techniques. The first thing to notice is that the topic level approach, using both PM-2 and xQuAD, significantly outperforms the baseline in all metrics, even with automatically generated topics. This confirms the effectiveness of both of these frameworks at providing results that are not only more relevant but also more diverse.

Interestingly, our term level approach also significantly outperforms the initial ranking in all measures. These results are consistent across both diversification techniques and topic sets. This confirms our intuition that identifying the right topic terms alone (without the grouping structure) is sufficient to improve both relevance and diversity with respect to the underlying topics. This has a very practical implication: it simplifies the task of generating a set of topics for each query to finding a set of terms.

This table also shows that the term-based approach maintains a comparable level of performance to the topic-based one. Their differences are not statistically significant across most of the measures. In fact, in the experiment with the ground-truth topics and terms, our approach is even slightly more robust: it helps more queries and hurts fewer. As we pointed out in Chapter 6, our approach may accidentally promotes documents for the "false" topics that correspond to the wrong combinations of terms. The fact that it achieves comparable performance to the topic level techniques indicates that this issue does have an impact on performance, but it is quite small.

Table 7.1: Performance comparison between term level ([Term]) diversification, topic level ([Topic]) and no diversification with respect to the baseline Query Likelihood (QL). Diversification is performed with respect to both the ground-truth topic sets (TREC) and the related queries (Related Q.) obtained from a commercial search engine. Win/Loss (W/L) is with respect to α-*NDCG*. † and ▼ indicate statistically significant differences to QL and the topic level approach respectively.

			Diversity							Relevance	
			CPR	α-NDCG	ERR-IA	NRBP	Prec-IA	S-Recall	W/L	NDCG	ERR
TREC		QL	0.5127	0.4156	0.3054	0.2679	0.1897	0.6215		0.2411	0.1446
	PM-2	[Topic]	$0.607^{\dagger}$	$0.5011^{\dagger}$	$0.3828^{\dagger}$	$0.3478^{\dagger}$	$0.2201^{\dagger}$	$0.6745^{\dagger}$	121/51	$0.2844^{\dagger}$	$0.1801^{\dagger}$
		[Term]	$0.5912^{\dagger}_{\blacktriangledown}$	$0.4847^{\dagger}_{\blacktriangledown}$	$0.3643^{\dagger}$	$0.3265^{\dagger}$	$0.2212^{\dagger}$	$0.6684^{\dagger}$	125/43	$0.2773^{\dagger}$	$0.1650^{\dagger}$
	xQuAD	[Topic]	$0.5993^{\dagger}$	$0.4936^{\dagger}$	$0.3743^{\dagger}$	$0.3395^{\dagger}$	$0.2208^{\dagger}$	$0.669^{\dagger}$	121/53	$0.2816^{\dagger}$	$0.1719^{\dagger}$
		[Term]	$0.5896^{\dagger}$	$0.4808^{\dagger}$	$0.3634^{\dagger}$	$0.327^{\dagger}$	$0.2226^{\dagger}$	$0.6587^{\dagger}$	122/47	$0.2808^{\dagger}$	$0.1668^{\dagger}$
Related Q.		QL	0.5076	0.4111	0.3025	0.2653	0.1844	0.6159		0.2401	0.1464
	PM-2	[Topic]	$0.5543^{\dagger}$	$0.4455^{\dagger}$	$0.3267^{\dagger}$	0.2857	$0.1982^{\dagger}$	$0.6509^{\dagger}$	100/66	$0.2597^{\dagger}$	0.152
		[Term]	$0.5534^{\dagger}$	$0.4466^{\dagger}$	$0.3301^{\dagger}$	$0.2914^{\dagger}$	0.1959	$0.6438^{\dagger}$	96/71	0.2544	0.1487
	xQuAD	[Topic]	$0.5486^{\dagger}$	$0.4402^{\dagger}$	0.3256	0.2861	$0.1976^{\dagger}$	$0.6389^{\dagger}$	91/73	0.25557	0.1482
		[Term]	$0.5537^{\dagger}$	$0.4462^{\dagger}$	$0.3336^{\dagger}$	$0.296^{\dagger}$	$0.2013^{\dagger}$	0.6355	96/67	$0.2649^{\dagger}$	0.1566

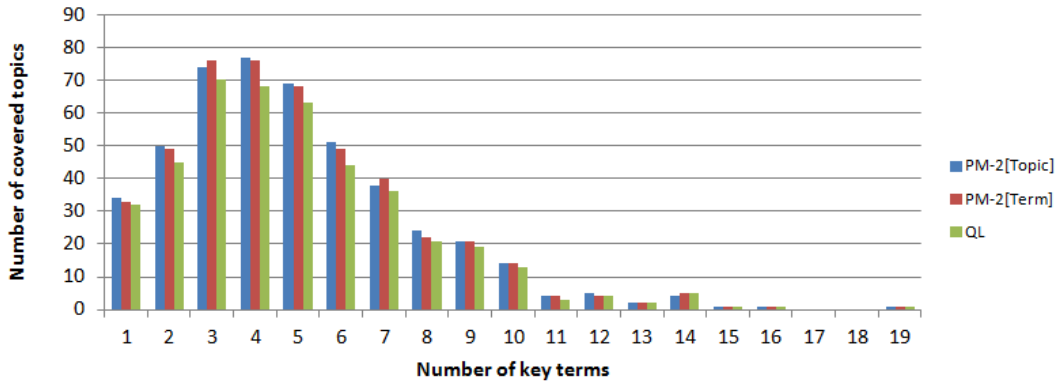
We notice, however, that some of the query topics are different from the query itself by only one term. For example, the topics for the query *"south africa"* include *"history of south africa"* and *"maps of south africa"*. Compared to the query, both of these topics only have one additional content-bearing term, which is *"history"* and *"maps"* respectively. Recall that our approach estimates the relevance of a document to a term by the geometric mean of the probability that this document generates this term and all of the query terms. For example, the relevance of d to *"history"* is estimated as follows:

$$P(d|history) \simeq (P\,(history|d) \times P\,(south|d) \times P\,(africa|d))^{\frac{1}{3}}$$

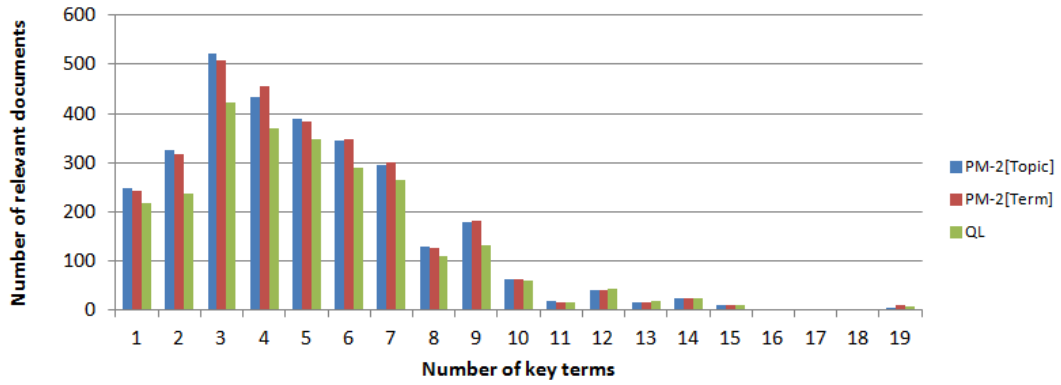
which is equivalent to the way the topic level approach estimates the relevance of this document d to the topic of *"history of south africa"*. As a result, it is possible that term level diversification is competitive with the topic level alternative because of queries like this. We are interested in seeing whether our approach can return relevant documents for the topics whose description is very different from the query.

To investigate this issue, we use the notion of *key term* to indicate the number of non-stopword terms in a query topic that are different from the query text. To quantify the impact the number of key terms has on our approach, we plot the number of topics where each approach is able to provide at least one relevant document against the number of key terms for these topics. In addition, we also plot the actual number of relevant documents retrieved for each topic against the number of key terms it contains. These plots are presented by Fig. 7.1 (a) and (b) respectively. Note that we only show the plots for `PM-2` because the analysis with `xQuAD` is very similar.

Fig. 7.1 reveals that that not only is our approach comparable with its topic counterpart on the topics with a single key term, it also remains competitive consistently across different numbers of key terms. In fact, our approach manages to cover more topics with 3 and 7 key terms than the topic-based system. This confirms the fact



(a) The total number of topics covered with respect to their number of key terms.



(b) The total number of relevant documents retrieved across all topics with respect to the number of key terms of these topics.

Figure 7.1: The total number of topics covered and the total number of relevant documents retrieved for all queries by each approach with respect to the number of key terms of each topic.

that PM-2 and xQuAD can take as input a set of terms and retrieve relevant documents for the topics underlying these terms.

In summary, our experiments show that PM-2 and xQuAD, although designed for topic level diversification, are capable of operating at the term level. In other words, they can take a set of topic terms and return documents for the latent topics underlying these terms. As a result, being able to identify an effective set of topic terms that span multiple topics is sufficient to improve diversity in the result rankings. The grouping structure indeed does provide additional benefits for diversification. However, given that such effective topical structures are very difficult to generate automatically, we argue that these benefits appear to be rather small.

7.2.2 SDM, RM and CA

Table 7.2 shows the effectiveness of our term level approach, with the ground-truth topic terms, when we use SDM, RM and CA to retrieve the initial ranking of documents. With SDM, this table shows a very similar trend to the previous results: our term level approach consistently outperforms the baseline in all relevance and diversity measures. This improvement is statistically significant in many cases. It is consistent across both diversification techniques. This further confirms that finding the right topic terms is sufficient to improve both relevance and diversity in the search results.

A part of this trend then carries over to the case with RM. Both PM-2 and xQuAD provide substantial and significant improvements in all of the diversity measures. Their improvement on the relevance measures, however, is not as consistent. While PM-2 and xQuAD achieve slightly higher *ERR* than RM, their *NDCG* is slightly lower. Regardless, these topic terms consistently improve diversity.

Table 7.2: Performance comparison between term level ([Term]) diversification, topic level ([Topic]) and no diversification with respect to three initial retrieval models: `SDM`, `RM` and `CA`. These results are with the ground-truth topics and terms. Win/Loss (W/L) is with respect to α-*NDCG*. † and ▼ indicate statistically significant differences to the initial models and the topic level approach respectively.

Ground-truth topics											
			Diversity							Relevance	
			CPR	α-NDCG	ERR-IA	NRBP	Prec-IA	S-Recall	W/L	NDCG	ERR
SDM		Base	0.5452	0.4393	0.329	0.2933	0.214	0.6204		0.2763	0.1578
	PM-2	[Topic]	$0.6137^{\dagger}$	$0.5094^{\dagger}$	$0.3878^{\dagger}$	$0.3525^{\dagger}$	$0.2329^{\dagger}$	$0.6787^{\dagger}$	113/63	$0.2983^{\dagger}$	$0.1858^{\dagger}$
		[Term]	$0.5907^{\dagger}_{\blacktriangledown}$	$0.4822^{\dagger}_{\blacktriangledown}$	$0.3603^{\dagger}_{\blacktriangledown}$	$0.3211^{\dagger}_{\blacktriangledown}$	$0.2263^{\dagger}_{\blacktriangledown}$	$0.6749^{\dagger}$	104/67	$0.2839_{\blacktriangledown}$	$0.1664_{\blacktriangledown}$
	xQuAD	[Topic]	$0.608^{\dagger}$	$0.5028^{\dagger}$	$0.3835^{\dagger}$	$0.3491^{\dagger}$	$0.2337^{\dagger}$	$0.6746^{\dagger}$	108/69	$0.2999^{\dagger}$	$0.1884^{\dagger}$
		[Term]	$0.59^{\dagger}_{\blacktriangledown}$	$0.4831^{\dagger}_{\blacktriangledown}$	$0.3647^{\dagger}$	$0.3278^{\dagger}$	$0.2265^{\dagger}$	$0.6675^{\dagger}$	104/69	$0.2832_{\blacktriangledown}$	$0.1684_{\blacktriangledown}$
RM		Base	0.5489	0.4406	0.3413	0.3106	0.221	0.5978		0.2911	0.1668
	PM-2	[Topic]	$0.6038^{\dagger}$	$0.5023^{\dagger}$	$0.3878^{\dagger}$	$0.3563^{\dagger}$	0.2299	$0.6669^{\dagger}$	116/57	0.3013	$0.1864^{\dagger}$
		[Term]	$0.5908^{\dagger}_{\blacktriangledown}$	$0.4832^{\dagger}_{\blacktriangledown}$	$0.3693^{\dagger}_{\blacktriangledown}$	$0.3349^{\dagger}_{\blacktriangledown}$	0.2263	$0.6565^{\dagger}$	101/69	$0.2899_{\blacktriangledown}$	$0.1699_{\blacktriangledown}$
	xQuAD	[Topic]	$0.5982^{\dagger}$	$0.49^{\dagger}$	$0.3767^{\dagger}$	$0.3438^{\dagger}$	0.2283	$0.6558^{\dagger}$	104/69	0.2938	0.1822
		[Term]	$0.589^{\dagger}$	$0.4774^{\dagger}$	$0.364^{\dagger}$	0.329	0.2273	$0.6492^{\dagger}$	99/71	0.287	0.1681
CA		Base	0.5881	0.4824	0.3719	0.3406	0.2457	0.6539		0.3146	0.1835
	PM-2	[Topic]	$0.632^{\dagger}$	$0.5257^{\dagger}$	$0.4089^{\dagger}$	$0.3763^{\dagger}$	0.2472	$0.6851^{\dagger}$	111/68	0.3191	0.1909
		[Term]	$0.6061_{\blacktriangledown}$	$0.4921_{\blacktriangledown}$	$0.3756_{\blacktriangledown}$	$0.3399_{\blacktriangledown}$	$0.2391_{\blacktriangledown}$	$0.6642_{\blacktriangledown}$	101/74	$0.3052_{\blacktriangledown}$	$0.1695_{\blacktriangledown}$
	xQuAD	[Topic]	$0.6274^{\dagger}$	$0.516^{\dagger}$	0.3979	0.3647	0.2448	$0.6779^{\dagger}$	101/75	0.3161	0.1931
		[Term]	$0.6001_{\blacktriangledown}$	$0.4895_{\blacktriangledown}$	$0.3736_{\blacktriangledown}$	$0.3377_{\blacktriangledown}$	$0.2356^{\dagger}_{\blacktriangledown}$	0.6642	101/73	$0.2965^{\dagger}_{\blacktriangledown}$	$0.1651_{\blacktriangledown}$

In the case of CA, we observe that PM-2 and xQuAD can still improve diversity with respect to most of the measures. Nevertheless, both of them hurt relevance. This can be explained by the fact CA models the relevance of a document to a query using about 100 features. The combination of these features that maximizes the relevance of the document ranking was learned from the training data in a supervised fashion. Our diversification systems, on the other hand, use Query Likelihood to determines the relevance of a document to each topic. It is thus reasonable that re-ordering the documents in a ranking that is highly optimized for relevance based on several query likelihood estimates result in a ranking with a lower degree relevance. The fact that PM-2 and xQuAD hurt relevance yet help diversity, in fact, provides strong evidence that the term level approach indeed can promote diversity with respect to query topics that underly the input set of terms. This is especially clear with the fact that both PM-2[Term] and xQuAD[Term] achieve higher *S-Recall* than CA.

Comparing our approach to the topic level technique, however, the difference in their performance is more significant in the case with SDM, RM and CA than with QL. This shows that correctly identifying the grouping structure between topic terms can indeed significantly improve performance. However, it is important to note that in practice, it is extremely difficult to generate topics that are as concise and accurate as these ground-truth topics. To the best of our knowledge, none of the existing techniques for topic generation have achieved this level of performance. In the next section, we will compare the effectiveness of the topics and terms that are generated automatically.

Regarding the topics and terms from the related queries, whose results are presented in Table 7.3, we observe a similar trend as well, although to a lesser extent. The improvement both approaches provide to the baseline is smaller and less consistent compared to using the ground-truth set. We believe that this is due to the mis-alignment issue where some of the related queries do not match up with the

ground-truth topics. The performance differences between the topic and term level techniques are also smaller.

In summary, our results have demonstrated that having the topic terms alone is sufficient for improving diversity effectiveness. This is consistent across two diversification techniques and four retrieval models for generating the initial document rankings. This shows that our term level approach to diversification is indeed effective.

7.3 Effectiveness of Generated Topic Terms

We will now study the practicality of our term level approach by investigating whether diversifying a document ranking with the topic terms that are generated automatically can improve diversity. For this purpose, we will use the topic terms generated using `DSPApprox` from the documents in the initial ranking. Specifically, we compare the effectiveness of these terms to that of the topics generated from the same set of documents using two existing techniques. This provides a comparison between our term level approach to the conventional topic level method in a practical setting where the ground-truth data is not available. After that, we examine the applicability of `PM-2` and `xQuAD`, two effective document diversification techniques, to the task of generating a diverse set of topic terms and compare them to `DSPApprox`.

7.3.1 `DSPApprox`

We employ `DSPApprox` to generate unigrams and phrases separately as topic terms. The total number of unigrams and phrases the algorithm returns are approximately 100 and 500 respectively. Since using too many terms is inefficient and unlikely to be effective, we use a parameter T to control the number of terms used for diversification.

Table 7.3: Performance comparison between term level ([Term]) diversification, topic level ([Topic]) and no diversification with respect to three initial retrieval models: SDM, RM and CA. The topics and terms are from the related queries provided by a commercial search engine. Win/Loss (W/L) is with respect to α-*NDCG*. † and ▼ indicate statistically significant differences to the initial models and the topic level approach respectively.

Related queries from a commercial search engine										
			Diversity							Relevance	
			CPR	α-NDCG	ERR-IA	NRBP	Prec-IA	S-Recall	W/L	NDCG	ERR
SDM		Base	0.5397	0.4329	0.3233	0.2874	0.2077	0.6161		0.2725	0.1582
	PM-2	[Topic]	$0.5746^{\dagger}$	$0.4587^{\dagger}$	$0.343^{\dagger}$	$0.3052^{\dagger}$	0.2128	$0.6417^{\dagger}$	97/64	0.2838	0.1697
		[Term]	$0.5675^{\dagger}_{\blacktriangledown}$	$0.4532^{\dagger}_{\blacktriangledown}$	0.3373	0.3001	0.2128	$0.6395^{\dagger}$	90/68	0.2825	0.1657
	xQuAD	[Topic]	$0.5666^{\dagger}$	0.4483	0.329	0.2884	0.2163	$0.642^{\dagger}$	92/69	0.281	0.1567
		[Term]	$0.5659^{\dagger}$	0.4512	0.3364	0.2969	0.2149	$0.6405^{\dagger}$	88/76	0.2776	0.1605
RM		Base	0.5424	0.4346	0.3368	0.3062	0.2128	0.5944		0.2851	0.1665
	PM-2	[Topic]	$0.5685^{\dagger}$	$0.4549^{\dagger}$	0.3428	0.3071	0.217	$0.6343^{\dagger}$	89/75	0.3013	$0.1864^{\dagger}$
		[Term]	$0.5587_{\blacktriangledown}$	0.4445	0.3341	0.2971	$0.211_{\blacktriangledown}$	$0.6268^{\dagger}$	82/79	$0.2899_{\blacktriangledown}$	$0.1699_{\blacktriangledown}$
	xQuAD	[Topic]	0.56	0.45	0.3404	0.3053	0.2146	$0.6302^{\dagger}$	90/67	0.2938	0.1822
		[Term]	0.5554	0.447	0.3381	0.3029	0.2128	$0.624^{\dagger}$	86/68	0.287	0.1681
CA		Base	0.5813	0.4753	0.3653	0.3336	0.2382	0.6508		0.3091	0.1832
	PM-2	[Topic]	0.5902	0.4849	0.3772	0.3469	0.2342	0.6548	102/63	0.3058	0.183
		[Term]	0.5853	0.4808	0.3745	0.3438	0.2362	0.647	97/68	0.3069	0.1827
	xQuAD	[Topic]	0.5812	0.478	0.371	0.3409	$0.2321^{\dagger}$	0.6453	98/65	0.3079	0.1854
		[Term]	0.5813	0.4779	0.3711	0.3413	$0.2322^{\dagger}$	0.6444	97/67	0.3073	0.1855

The second parameter is w, which determines the size of the window in which (1) a term has to co-occur with at least one query term in order to be considered a candidate topic term, and (2) prediction boundary: a term cannot predict terms that are more than w words away. We set $w = 20$ based on our prior experiments (Dang & Croft, 2013) and tune $T \in \{5, 10, 20, 40, 60, 80, 100\}$ via 5-fold cross-validation.

We consider three baselines for comparison.

Baseline 1. Our first baseline was the technique proposed by Carterette and Chandar (2009). It applies `LDA` (Blei et al., 2003) on the initially retrieved documents and uses the resulting LDA topics for diversification. The process that generates these topics also provides an estimate of how relevant each document is to each of these topics, which can be used as the $P(d|t)$ component in both `PM-2` and `xQuAD`. Each LDA topic is essentially a distribution of terms. Another way to estimate $P(d|t)$ is thus to treat this distribution as a weighted query and set $P(d|t)$ to the query likelihood score of the document with respect to this query. We will report the results using the latter estimate of $P(d|t)$ since we found it to be more effective. This technique has two parameters that need tuning. The first parameter is the number of latent topics c. We consider $c \in \{2, 5, 10\}$. The second parameter is the number of the most highly weighted terms T from each topic we use to form the weighted query. We consider $w \in \{5, 10, 50\}$. We use the multi-threaded implementation of LDA that is publicly available [1].

Baseline 2. Our second baseline technique, also proposed by Carterette and Chandar (2009), applies k-nearest neighbor (`KNN`) to cluster the retrieved documents. After that, it estimates a relevance model (Lavrenko & Croft, 2001) from each of the clusters and use it as a topic model. Similarly, $P(d|t)$ is estimated as the query likelihood score of the document to the weighted query constructed from the T most highly

[1] https://sites.google.com/site/rameshnallapati/software

weighted terms from each relevance model. Its parameters include $k \in \{2, 5, 10\}$ and $T \in \{5, 10, 20\}$, which are the number of nearest neighbors and the number of top terms from the relevance model to be used to form the query respectively.

Baseline 3. `MMR` (Carbonell & Goldstein, 1998) has become a canonical baseline in the diversity literature. Though it does not explicitly model topics, it fits into the class of algorithms that relies solely on the set of documents. It greedily selects documents from the initial ranking with the following objective function:

$$d^* \leftarrow \arg\max_{d \in R} \lambda R(d, q) - (1 - \lambda) \max_{d_j \in S} Sim(d, d_j)$$

where $R(d, q)$ is the relevance of the document d to the query q which can be acquired directly from the baseline ranking and $Sim(d, d_j)$ is the cosine similarity between the two documents.

All parameters associated with the three baselines are determined using 5-fold cross validation. Topics and terms are extracted from the top 100 documents in the initial rankings. We first present our results and analyses for the case where `QL` is used to retrieve the initial rankings. After that, we will discuss the case with `SDM`, `RM`, and `CA`.

7.3.1.1 Results with Query Likelihood (`QL`)

Table 7.4 presents the comparison between the techniques mentioned above. The letters b, m, k and l indicate statistically significant differences (p-value < 0.05) to query-likelihood, `MMR`, `KNN` and `LDA` respectively. Among the techniques under investigation, `MMR` hurts more queries than it helps and thus fails to provide any improvement over the baseline. The two topic level techniques, `LDA` and `KNN`, using either `PM-2` or `xQuAD` for diversification, increase performance in some measures (e.g. *NRBP*) but decrease the performance in other measures (e.g. α-*NDCG* and *CPR*). They fail to provide consistent improvement overall. Furthermore, the difference between their

performance and that of the initial ranking is very small and not statistically significant.

In contrast, both the unigrams and phrases generated using DSPApprox, when used by both PM-2 and xQuAD, substantially outperform all other systems under comparison in almost all measures. Statistically significant differences are observed in many cases. This confirms the effectiveness of our term level approach as well as DSPApprox at generating very effective sets of topic terms. Between unigrams and phrases, the former appears to be slightly more robust by improving more queries and hurting fewer, but the latter manages to retrieve more relevant results.

Table 7.5 presents two example topics provided by LDA and KNN and the terms provided by DSP[U] for the query *"sat"*. Two of the ground-truth topics for this query include *typical good range of SAT scores* and *information on test preparation materials and courses for sat*. We examined the top 30 terms from these topics and manually selected all of the terms that we believe are useful. For readability, we present these terms in bold and before the less useful ones. The bolded terms for DSP[U] are selected in the same manner. This table shows the first topic provided by KNN and LDA is indeed very reasonable. The second one, however, is not. This is the general problem with clustering where the number of clusters is specified in advance. The non-relevant documents in the initial ranking usually produce one or more non-relevant clusters, which hurts diversification. Furthermore, even the topics with the effective terms contain unrelated words. This might also have a negative impact on diversity.

Table 7.4: Performance comparison among approaches that use PM-2 and xQuAD for diversification with respect to (1) topic terms (both unigrams and phrases) generated by DSPApprox (abbreviated as DSP[U] and DSP[P]) and (2) topics generated by LDA and KNN. The baseline retrieval model is query-likelihood (QL). In addition, we also compare their results with MMR, which does not explicit model query topics. Evaluation is done using a wide range of diversity and relevance measures. Win/Loss (W/L) is with respect to α-*NDCG*. b, m, k and l indicate statistically significant differences (p-value < 0.05) to the baseline QL, MMR, KNN and LDA respectively. Bold face indicates the best performance in each group.

			Diversity							Relevance	
			CPR	α-NDCG	ERR-IA	NRBP	Prec-IA	S-Recall	W/L	NDCG	ERR
QL		Base	0.5127	0.4156	0.3054	0.2679	0.1897	0.6215		0.2411	0.1446
		MMR	0.5077_{b}	0.4109_{b}	0.3018_{b}	0.2641	0.1882	0.614	55/85	0.2369_{b}	0.1408_{b}
	PM-2	KNN	0.5115	0.412	0.3062	0.272	0.1926	0.605	81/81	0.2483	0.1506_{m}
		LDA	0.5152	0.4143	0.3064	0.2713	0.1996_{m}	0.6061	86/78	0.2561_{m}	0.1516_{m}
		DSP[U]	$\mathbf{0.5456}^{k,l}_{b,m}$	$\mathbf{0.4426}^{k,l}_{b,m}$	$\mathbf{0.3343}^{k,l}_{b,m}$	$\mathbf{0.3003}^{k,l}_{b,m}$	$\mathbf{0.208}^{k,l}_{b,m}$	**0.6238**	100/65	$0.2671^{k,l}_{b,m}$	$0.1606^{k,l}_{b,m}$
		DSP[P]	0.5304	0.4317^{k}_{m}	$0.3294^{k,l}_{b,m}$	$0.2995^{k,l}_{b,m}$	$\mathbf{0.208}^{k}_{b,m}$	0.6048	90/80	$\mathbf{0.2682}^{k}_{b,m}$	$\mathbf{0.1646}^{l}_{b,m}$
	xQuAD	KNN	0.5143_{m}	0.4143	0.3049	0.2685	0.1904	0.6161	56/43	0.2427_{m}	0.1447_{m}
		LDA	0.513_{m}	0.4151	0.3044	0.2668	0.1909	**0.6225**	44/43	0.2419_{m}	0.1445_{m}
		DSP[U]	$\mathbf{0.5397}^{k,l}_{b,m}$	$\mathbf{0.4364}^{k,l}_{b,m}$	$\mathbf{0.3293}^{k,l}_{b,m}$	$0.2964^{k,l}_{b,m}$	$0.2085^{k,l}_{b,m}$	0.6179	97/67	$0.2665^{k,l}_{b,m}$	$0.1592^{k,l}_{b,m}$
		DSP[P]	0.5282	0.4278	$0.3275^{k,l}_{m}$	$\mathbf{0.2979}^{k,l}_{b,m}$	$\mathbf{0.2094}^{k,l}_{b,m}$	0.5914^{l}_{b}	86/83	$\mathbf{0.2702}^{k,l}_{b,m}$	$\mathbf{0.1669}^{k,l}_{b,m}$

Table 7.5: Topics extracted by KNN and LDA and terms extracted by DSPApprox for the query *"sat"*. Two of the ground-truth topics of this query are *typical good range of SAT scores* and *information on test preparation materials and courses for sat.*

KNN		LDA		DSP[U]	
Topic 1	Topic 2	Topic 1	Topic 2	Terms	
test	vegan	**test**	pm	**college**	take
prep	potluck	**vocabulary**	jazz	**test**	july
score	6pm	**prep**	open	**vocabulary**	potluck
guide	10pm	**score**	fri	**prep**	vegan
preparation	northridge	**preparation**	2006	**book**	satisfy
college	tour	satisfy	rev	**course**	april
grenadines	dec	dissatisfy	aug	**exam**	june
vincent	los	answer	thu	**essay**	february
mahalo	2009	question	tue	**min**	march
grenada	fri	neutral	karaoke	**score**	january

The terms generated by DSP[U] provides good coverage for the first ground-truth topic and might slightly cover the second (*"min"* and *"score"*). As we have demonstrated earlier, PM-2 and xQuAD are able to take a set of topic terms as input and return documents for the underlying topics, each of which corresponds to a certain combination of those terms. The fact that DSPApprox significantly outperforms both the baseline ranking and the two topic-based systems further supports this claim as well as confirming the effectiveness of DSPApprox.

Although DSPApprox also generates off-topic terms, the superiority of both DSP[U] and DSP[P] suggests that the effect of these terms on our approach is not as significant as it is on the topic level alternative. This may be due to the fact that they are not clustered with the on-topic terms, thus they do not interfere as much with these terms in representing the underlying topics. We will study their effect in Section 7.3.1.3.

7.3.1.2 Improvement Analysis

We focus our analysis on the DSPApprox results using PM-2. Our findings apply to xQuAD as well. As can be seen from Table 7.4, DSPApprox significantly improves

the baseline ranking in all diversity measures except *S-Recall*. Specifically, `DSP[U]` increases *S-Recall* slightly while `DSP[P]` decreases it. Our investigation suggests that not only do `DSP[U]` and `QL` cover about the same number of topics, they cover almost the same set of topics (98% overlap). This indicates that the improvement made by `DSPApprox` comes from better per topic coverage, which refers to both more relevant documents for each of the covered topics as well as better ranking of these documents. Quantitative analysis of the effect of better per topic coverage on performance is provided in Table 7.6. WIN and LOSS indicate the set of queries where `DSPApprox` helps and hurts α-*NDCG* compared to `QL`. $\%\Delta$P denotes the relative performance difference in α-*NDCG*. *S.Rec* $\uparrow$ indicates the subset of WIN where S-Recall is also improved and *REST* indicates the remaining of the set. Similarly, *S.Rec* $\downarrow$ indicates the subset of LOSS where S-Recall is also lower and *REST* indicates the remaining. The performance difference between `DSPApprox` and `QL` in the REST set indicates the contribution of better within topic coverage to the overall improvement brought by `DSPApprox`. It can be seen that 62.14% and 69.96% of the improvement provided by `DSP[U]` and `DSP[P]` respectively comes from having better within topic coverage. Similarly, 57.06% and 53.89% of the decrease in α-*NDCG* caused by these two systems are also attributed to per topic coverage. Since the percentage decrease in α-*NDCG* is significantly lower than the percentage improved, both `DSP[U]` and `DSP[P]` outperform `QL` overall. It is important to point out the performance differences in the *S.Rec* $\uparrow$ and *S.Rec* $\downarrow$ segments do not indicate the effect of having broader topic coverage alone but also that of per topic coverage.

Table 7.6 also demonstrates that `DSP[U]` improves and hurts *S-Recall* for roughly the same number of queries, which accounts for the slight overall improvement in *S-Recall*. `DSP[P]`, on the other hand, hurts more queries than it helps with respect to *S-Recall*. This explains the overall decrease in this measure.

Table 7.6: Contribution of within topic coverage to the overall improvement in α-*NDCG*. Within topic coverage refers to both having more relevant documents for each of the covered topics and better ranking of these documents. WIN and LOSS indicate the sets of queries whose α-*NDCG* `DSPApprox` (abbreviated as `DSP`) improves and hurts respectively. $S.Rec \uparrow$ is the subset of WIN on which subtopic recall is also improved and *REST* is its complement. $S.Rec \downarrow$ is the subset of LOSS on which subtopic recall is also lowered and *REST* is its complement. ΔP is the relative difference in α-*NDCG* between `DSPApprox` and `QL`. `[U]` and `[P]` indicate terms and phrases respectively.

		%ΔP		#q	Contribution to %ΔP
`DSP[U]`	WIN	+19.54%	$S.Rec \uparrow$	16	+37.86%
			REST	84	+**62.14**%
	LOST	−11.25%	$S.Rec \downarrow$	14	−42.94%
			REST	51	−**57.06**%
`DSP[P]`	WIN	+26.27%	$S.Rec \uparrow$	16	+30.04%
			REST	74	+**69.96**%
	LOST	−16.76%	$S.Rec \downarrow$	23	−46,11%
			REST	57	−**53.89**%

The analyses above suggest that the terms provided by `DSPApprox`, though unable to provide much broader coverage for the query topics, correctly represent most of those covered by `QL`. Consequently, they help surface more documents on these topics, significantly improving diversity according to *CPR*, all three cascade measures, *Precision-IA*, as well as both relevance metrics.

The fact that diversification with both unigrams and phrases given by `DSPApprox` significantly improves the relevance of the results (*NDCG* and *ERR*) is very interesting. Our approach, in fact, is very similar to pseudo-relevance feedback. The difference is that traditional relevance feedback uses the extracted terms to update the query model to retrieve new documents. Our approach, on the other hand, only attempts to re-order the input ranking, pushing more relevant documents to earlier ranks. As such, diversification can be considered a precision-driven framework for relevance feedback.

7.3.1.3 Failure Analysis

As mentioned above, `DSP[U]` does not provide significant improvement in *S-Recall*. In fact, `DSP[P]` hurts *S-Recall*. Our analysis shows that `DSP[U]` and `QL` cover almost the same set of topics (98% overlap). This high percentage of overlap suggests that the terms generated by `DSPApprox` are biased towards topics covered by the top ranked documents in the initial ranking.

We believe the cause of this bias is the way `DSPApprox` computes topicality. We observe that the topicality of a term is relatively proportional to the relevance model probability (Lavrenko & Croft, 2001) estimated from the initially retrieved documents. This model usually assigns higher probabilities to frequent terms from higher ranked documents since they are assumed to be more relevant. If a document at a very low position covers topics that are different from those at early ranks, chances are their topic terms do not appear in these documents with high frequency. Therefore, their chance to be included in the resulting set of terms is relatively small, causing these topics to be excluded from the coverage of the final set. This is the main reason why subtopic recall was not improved.

We have also found two other causes of failure which are due to the combined effect of several factors. First, there is the "false topic" issue due to arbitrary combinations of the extracted terms that we have discussed earlier. This is a general problem with our term level approach. Table 7.7 shows some example terms extracted by `DSPApprox` for the query *"kenmore gas water heater"*. It turns out that the documents that mention gas water heaters manufactured by Kenmore also mention a variety of other electric appliances. In addition to these relevant documents, Query Likelihood also returns several non-relevant documents about water heater from other manufacturers and other appliances by Kenmore. As a result, while `DSPApprox` found some good terms that align effectively with the ground-truth topics such as *"manual"*, *"tankless"* and *"steam"* (see Table 7.7), it also returns off-topic terms such as

"electric", *"appliances"* and *"refrigerators"*. One of the possible "false topics" that arises is *"kenmore refrigerators"*. When PM-2 attempts to diversify this ranking with respect to those terms, it promotes the non-relevant documents that mention both *"kenmore refrigerators"* and *"aquastar gas water heater"*. The reason is that the former phrase makes this non-relevant document appear highly relevant to the "false topic" while the latter makes this non-relevant document appear somewhat relevant to the query.

Table 7.7: Some example outputs of DSPApprox for the query *"kenmore gas water heater"*. Important terms from the original TREC subtopics for this query are also provided.

TREC Sub-topic	DSPApprox
1) reviews	
2) owner manuals	manual
3) features, energy consumption and safety	tankless
	steam
	electric
	appliances
	refrigerators

The second cause of failure is due to the fact that the non-relevant documents retrieved by Query Likelihood are consistently on the wrong topics. For example, for the query *"adobe indian house"*, in addition to the relevant documents, QL also returned quite a few documents about *"adobe pdf"*. Although DSPApprox manages to extract good terms from the relevant documents, the non-relevant ones cause DSPApprox to generate terms such as *"acrobat"* and *"pdf"*. As a result, PM-2 promotes documents for this non-relevant topic. Example terms are presented in Table 7.8.

7.3.1.4 Results with SDM, RM and CA

Table 7.9 and Table 7.10 compare the term level approach and topic level alternative with SDM, RM and CA as the baseline retrieval models. Overall, although the performance difference between systems is smaller, the trend is similar to the case with QL as the initial retrieval model. Diversification using the set of topic terms generated

Table 7.8: Some example outputs of DSPApprox for the query *"adobe indian house"*. Important terms from the original TREC subtopics for this query are also provided.

TREC Sub-topic	DSPApprox
1) How to build	build
2) Indian tribes that used adobe houses	pueblo
	tribe
3) books, videos about adobe building	
	acrobat
	pdf

by DSPApprox can improve both relevance and diversity in most cases compared the baseline ranking regardless of what model being used.

Between the topic level and our term level approach, our approach with PM-2 as the diversification technique (both DSP[U] and DSP[P]) consistently outperforms both KNN and LDA in both *CPR* and all three cascade diversity measures. Although this improvement is only statistically significant in a few cases, it is consistent across SDM, RM and CA. In addition, DSP[U] consistently achieves higher *S-Recall* than the two topic level methods. This further supports our claims. Firstly, having the right set of topic terms, which we can generate automatically, is sufficient to improve diversity with respect to the underlying topics. Secondly, it is very difficult to generate accurate topic structures that are beneficial to diversification. A similar trend can be seen with xQuAD, although it is less consistent.

The benefits of using phrases appear to be larger with more effective initial rankings. Both with SDM and CA, DSP[P] using either PM-2 or xQuAD is the top performing approach with respect to all three cascade diversity measures as well as the relevance measures.

Table 7.9: Performance comparison among systems that use PM-2 and xQuAD for diversification with respect to (1) topic terms (both unigrams and phrases) generated by DSPApprox (abbreviated as DSP[U] and DSP[P]) and (2) topics generated by LDA and KNN. We consider two baseline retrieval models: SDM and RM. In addition, we also compare their results with MMR. Win/Loss is with respect to α-$NDCG$. b, m, k and l indicate statistically significant differences (p-value < 0.05) to the baseline Base, MMR, KNN and LDA respectively. Bold face indicates the best performance.

			Diversity							Relevance	
			CPR	α-NDCG	ERR-IA	NRBP	Prec-IA	S-Recall	W/L	NDCG	ERR
SDM		Base	0.5452	0.4393	0.329	0.2933	0.214	0.6204		0.2763	0.1578
		MMR	0.5382_{b}	0.4343_{b}	0.3226_{b}	0.2847_{b}	0.2106_{b}	0.6272	43/105	0.2711_{b}	0.1527_{b}
	PM-2	KNN	0.547	0.4405	0.3363_{m}	0.3035_{m}	0.2148	0.611	92/72	0.2811	$0.1672_{b,m}$
		LDA	0.5353^{k}	0.4358	0.3306	0.2989	0.2112	0.6099	78/86	0.2728^{k}	0.163_{m}
		DSP[U]	$\mathbf{0.5614}^{k,l}_{b,m}$	0.451^{l}_{m}	0.3423_{m}	0.3093_{m}	0.2172	$\mathbf{0.6371}^{l}$	89/79	0.2756	0.1642
		DSP[P]	0.5595^{l}_{m}	$\mathbf{0.4517}^{l}_{m}$	$\mathbf{0.3484}^{l}_{b,m}$	$\mathbf{0.3179}^{l}_{b,m}$	$\mathbf{0.2313}^{k,l}_{b,m}$	0.6107	78/90	$\mathbf{0.3029}^{k,l}_{b,m}$	$\mathbf{0.1777}^{l}_{b,m}$
	xQuAD	KNN	0.5454_{m}	0.4376	0.3294_{m}	0.2942_{m}	0.2153_{m}	0.6174	67/51	0.2771_{m}	0.1618_{m}
		LDA	0.5454_{m}	0.4387_{m}	0.3303_{m}	0.2953_{m}	0.2139_{m}	0.617_{m}	60/46	0.2756_{m}	0.1586
		DSP[U]	0.5446	0.4367	0.3282	0.2936	0.2189	**0.6265**	90/77	0.2732	0.1599
		DSP[P]	**0.5467**	**0.4446**	**0.3419**	$\mathbf{0.3105}_{m}$	$\mathbf{0.2324}^{k,l}_{b,m}$	0.6173	86/83	$\mathbf{0.298}^{k,l}_{b,m}$	$\mathbf{0.1747}_{m}$
RM		Base	0.5489	0.4406	0.3413	0.3106	0.221	0.5978		0.2911	0.1668
		MMR	0.5493	0.4407	0.3399	0.3083	0.2206	0.6035	61/79	0.2874_{b}	0.1623_{b}
	PM-2	KNN	0.5467	0.4426	0.3443	0.314	0.2215	0.5999	60/76	0.2904	0.1717_{m}
		LDA	0.537_{m}	0.4339	0.3396	0.312	$0.2134^{k}_{b,m}$	0.5896_{m}	62/85	0.2831^{k}	0.1678
		DSP[U]	$\mathbf{0.5593}^{k,l}$	$\mathbf{0.4512}^{l}_{b,m}$	0.3475	0.317	$\mathbf{0.2254}^{l}$	$\mathbf{0.6127}^{l}$	88/75	0.2907	0.1697
		DSP[P]	0.5425	0.4482^{l}	**0.3502**	**0.3218**	0.2247^{l}	0.6046	87/68	$\mathbf{0.2969}^{l}$	**0.1726**
	xQuAD	KNN	0.5491	0.4399	0.3404	0.3093	0.2227	0.5974	38/42	0.2913_{m}	0.1665_{m}
		LDA	0.5481	0.442	0.3432	0.3131	0.222	0.5991	39/61	0.2913_{m}	$\mathbf{0.1709}_{m}$
		DSP[U]	**0.5523**	$\mathbf{0.4514}^{k}_{b,m}$	**0.349**	**0.3189**	0.2242	$\mathbf{0.6151}^{k,l}_{b}$	85/68	0.2908	0.1669
		DSP[P]	0.5466	0.4446	0.3446	0.3145	**0.2257**	0.6083	78/77	$\mathbf{0.3004}^{k,l}_{b,m}$	0.1707

Table 7.10: Performance comparison among systems that use PM-2 and xQuAD for diversification with respect to (1) topic terms (both unigrams and phrases) generated by DSPApprox (abbreviated as DSP[U] and DSP[P]) and (2) topics generated by LDA and KNN. The baseline retrieval model is CA. In addition, we also compare their results with MMR. Win/Loss is with respect to α-*NDCG*. b, m, k and l indicate statistically significant differences (p-value < 0.05) to the baseline Base, MMR, KNN and LDA respectively. Bold face indicates the best performance.

			Diversity							Relevance	
			CPR	α-NDCG	ERR-IA	NRBP	Prec-IA	S-Recall	W/L	NDCG	ERR
CA		Base	0.5881	0.4824	0.3719	0.3406	**0.2457**	0.6539		0.3146	0.1835
		MMR	0.5882	0.4826	0.372	0.3407	0.2449_{b}	**0.6552**	41/46	0.3141	0.1835
	PM-2	KNN	0.5847	0.4871	0.385	0.3566	$0.2402_{b,m}$	0.6399_{m}	94/75	0.3163	**0.1932**
		LDA	0.584	0.4866	0.3841	0.3556	$0.2397_{b,m}$	0.6419	89/79	0.3156	0.19
		DSP[U]	0.5876	0.4924	0.3902	0.3624	0.2406_{b}	0.6437	97/72	0.317	0.1919
		DSP[P]	$\mathbf{0.5903}^{l}$	$\mathbf{0.4937}^{l}$	$\mathbf{0.3924}^{l}_{b,m}$	$\mathbf{0.3652}^{l}_{b,m}$	0.2412	$0.6385_{b,m}$	98/73	**0.319**	0.1902
	xQuAD	KNN	0.5848	0.4833	0.3768	0.3474	$0.239_{b,m}$	0.645	98/71	0.3118	0.1856
		LDA	0.5835	0.4835	0.3776	0.3484	$0.2387_{b,m}$	0.6437	95/75	0.3129	0.1887
		DSP[U]	0.5844	0.4824	0.376	0.3464	$0.2388_{b,m}$	0.6463	99/71	0.3115	0.1842
		DSP[P]	0.587^{l}	**0.4891**	$\mathbf{0.3856}^{l}$	$\mathbf{0.3577}^{k,l}$	$0.2403_{b,m}$	0.6439	97/73	**0.3157**	**0.1898**

7.3.2 Term Generation with Document Diversification Methods

We now evaluate the effectiveness of PM-2 and xQuAD for generating a diverse set of topic terms. Recall that this term generation task is also a diversification problem. Instead of generating a ranking of documents with coverage for multiple topics, the aim is to select a set of topic terms that can predict the occurrence of a large proportion of the vocabulary associated with some set of latent topics. Consequently, PM-2 and xQuAD can be applied straight-forwardly, as described in Chapter 6. To ensure fair comparison between PM-2, xQuAD and DSPApprox, we set $\lambda = 0.5$ for both PM-2 and xQuAD.

The comparison is done as follows. We first use DSPApprox to extract a set of unigram topic terms from the initial ranking of documents as we did earlier. We then provide the same set of vocabulary $v \in V$, candidate topic terms $t \in T$, their topicality $TP(t)$ and the predictiveness statistics $P(t|v)$ to PM-2 and xQuAD. These techniques will then generate their own set of topic terms. Each of these sets are then provided to PM-2 to diversify the initial ranking of documents retrieved by each of the four standard models: QL, SDM, RM and CA.

The results are presented in Table 7.11. It shows that although the terms generated using PM-2 and xQuAD can provide some improvement over the initial rankings, they are not consistent. Furthermore, they are not as effective as those provided by DSPApprox. Recall that the difference between these techniques is as follows. After a topic term is selected, PM-2 and xQuAD downweight the vocabulary words that are covered (or predicted) by this term by some amount specified as part of their objective function. In the next iteration, the value of those remaining topic terms that predict these same vocabulary will thus be discounted. DSPApprox, on the other hand, completely disregards these vocabulary words in future iterations. As a result, DSPApprox can be considered more aggressive in the treatment of vocabulary that has been somewhat covered.

Let us revisit the objective function of `PM-2` ($\lambda = 0.5$):

$$t^* \leftarrow \arg\max_{t \in T} \sum_{v \in V_t} q_v \times TP(t) \times P(t|v)$$

where q_v is the quotient of the vocabulary v (i.e., its weight), which is computed as:

$$q_v = \frac{1}{2\left(\sum_{t \in S} \frac{P(t|v)}{\sum_{v' \in V} P(t|v')}\right) + 1}$$

where S is the set of selected topic terms at the current iteration. After a topic term is selected (i.e., added to S) that covers a vocabulary v, the quotient q_v will become lower in the next iteration. However, given that our vocabulary set typically consists of thousands of words, the amount of quotient decreased from one iteration to the next for v is very small. As a result, in the event that most of the vocabulary is from one latent topic while less is from another topic, it is possible that `PM-2` keeps selecting the topic terms that predict the vocabulary of this dominating topic, leaving the vocabulary of the other topic uncovered. One can verify that this can also happen to `xQuAD`. We argue that this is why `PM-2` and `xQuAD` do not work as well as `DSPApprox`.

To verify if this is true, we introduce two variants of `PM-2` for term generation, which involves a more aggressive downweighting function: `PM-2L` (linear discounting) and `PM-2E` (exponential discounting). Their formula for computing the quotient is as follows:

$$q_v^{(PM-2L)} = \frac{1}{2|S_v|\left(\sum_{t \in S} \frac{P(t|v)}{\sum_{v' \in V} P(t|v')}\right) + 1}$$

$$q_v^{(PM-2E)} = \frac{1}{2e^{|S_v|}\left(\sum_{t \in S} \frac{P(t|v)}{\sum_{v' \in V} P(t|v')}\right) + 1}$$

where $|S_v|$ is the number of topic terms in S such that $P(t|v) > 0$ (i.e., t covers v to some extent). This is based on the fact that `DSPApprox` disregards a vocabulary word v in future iterations as soon as a topic term t with $P(t|v) > 0$ is selected.

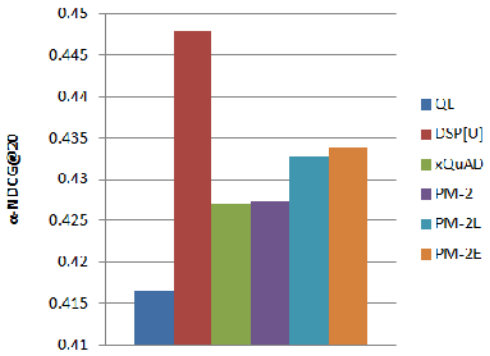
Figure 7.2 compares the terms provided by PM-2L and PM-2E with the original PM-2. We also show the results with the terms generated using DSPApprox and xQuAD for comparison. They are evaluated based on their effectiveness for document diversification measured using α-$NDCG$. It is clear that with respect to PM-2, having a more aggressive downweighting function increases the diversity among the selected topic terms. It also appears exponential downweighting is more effective overall. Regardless, both PM-2L and PM-2E are still not as good as DSPApprox, which ignores a vocabulary as soon as it has been predicted instead of downweighting it.

7.4 Results with Terms Extracted from External Sources

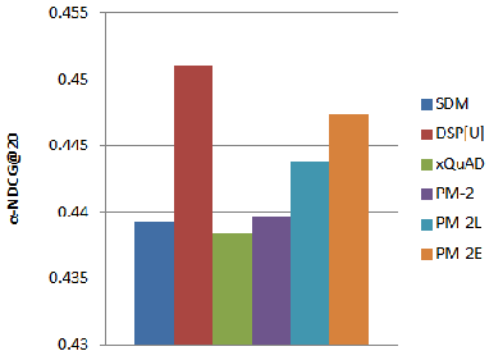
In this section, we evaluate the effectiveness of the five external sources of information, the AOL log (AOL), the MSN log (MSN), anchor text (Anchor), Freebase (Freebase) and Wikipedia (Wiki), for providing useful terms that are beneficial to diversification. In addition to using the topic terms from each resource separately, we test a system that combines the top T terms from each of the four resources (All). We will also compare their effectiveness with that of the retrieved documents (RDOC), which we have discussed in the previous section. Recall that with the external resources, we only retrieve the "documents" (which are queries in the case with query logs, entity descriptions in the case with Freebase, etc.) that contain all of the query terms. As a result, each resource does not provide documents for all 200 queries. For comparison, we use the subset of queries where all of them can provide at least two documents, on which we can apply DSPApprox to obtain the topic terms. The resulting subset consists of 98 queries. The parameter T, which is the number of topic terms used by each system under comparison, as well as all parameters of the diversification techniques being used are tuned via 5-fold cross-validation.

Table 7.11: Effectiveness for diversification of the topic terms generated by DSPApprox, xQuAD and PM-2. Document diversification is done using PM-2 on top of an initial document ranking retrieved by QL, SDM, RM and CA. Win/Loss is with respect to α-*NDCG*. † and ▼ indicate statistically significant differences (p-value < 0.05) to the baseline Base and DSPApprox respectively. Bold face indicates the best performance.

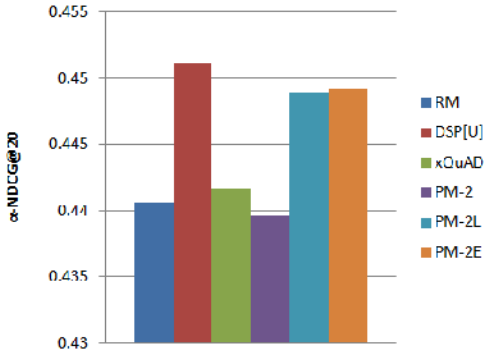
		Diversity							Relevance	
		CPR	α-NDCG	ERR-IA	NRBP	Prec-IA	S-Recall	W/L	NDCG	ERR
QL	Base	0.5132	0.4165	0.306	0.2688	0.1872	0.6234		0.2239	0.139
	DSP[U]	**0.5514**†	**0.448**†	**0.3383**†	**0.3044**†	**0.2093**†	**0.633**	90/56	**0.2538**†	**0.1568**†
	PM-2	0.5328†	0.4273$_{\blacktriangledown}$	0.3224$^{\dagger}_{\blacktriangledown}$	0.2905$^{\dagger}_{\blacktriangledown}$	0.2027$_{\blacktriangledown}$	0.6139	81/67	0.2452†	0.1535†
	xQuAD	0.5284†	0.4271$_{\blacktriangledown}$	0.3173$_{\blacktriangledown}$	0.2818$_{\blacktriangledown}$	0.1977$_{\blacktriangledown}$	0.6276	81/70	0.2376†	0.1446†
SDM	Base	0.5452	0.4393	0.329	0.2933	0.214	0.6204		0.2763	0.1578
	DSP[U]	**0.5614**†	**0.451**	**0.3423**	**0.3093**	0.2172	**0.6371**	89/79	0.2756	**0.1642**
	PM-2	0.5522	0.4396$_{\blacktriangledown}$	0.3334	0.3012	0.2178	0.6085$_{\blacktriangledown}$	84/78	**0.2798**	0.1599
	xQuAD	0.5463†	0.4384$_{\blacktriangledown}$	0.3285	0.2928	0.2154	0.6211	87/75	0.2757	0.1575
RM	Base	0.5489	0.4406	0.3413	0.3106	0.221	0.5978		**0.2911**	0.1668
	DSP[U]	**0.5593**	**0.4512**	**0.3475**	**0.317**	**0.2254**	**0.6127**	88/75	0.2907	**0.1697**
	PM-2	0.5473†	0.4396$_{\blacktriangledown}$	0.339	0.31	0.2166$_{\blacktriangledown}$	0.5962$_{\blacktriangledown}$	76/79	0.2795†	0.1576†
	xQuAD	0.5493	0.4417$_{\blacktriangledown}$	0.3383	0.3075	0.2176$_{\blacktriangledown}$	0.6089	82/72	0.2805†	0.1577†
CA	Base	0.5881	0.4824	0.3719	0.3406	0.2457	0.6539		0.3146	0.1835
	DSP[U]	0.5876	**0.4924**	**0.3902**	**0.3624**	0.2406†	0.6437	97/72	**0.317**	**0.1919**
	PM-2	0.5858	0.488	0.3837	0.3547	0.238†	0.6448	94/74	0.3111	0.1885
	xQuAD	**0.5883**	0.4877	0.3843	0.355	0.2399†	0.6386†	96/74	0.3137	0.1896



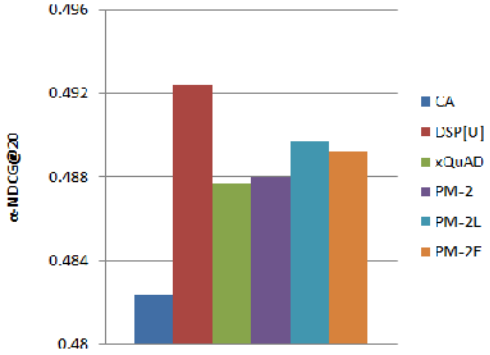
(a) Initial rankings retrieved using QL



(b) Initial rankings retrieved using SDM.



(c) Initial rankings retrieved using RM.



(d) Initial rankings retrieved using CA.

Figure 7.2: Comparison of diversification effectiveness among the topic terms provided by PM-2 and its two variants with more aggressive down-weighting strategy: PM-2L and PM-2E. In addition, we also show the performance of the terms extracted using DSPApprox and xQuAD as well as the initial rankings for comparison. It can be seen that discounting the vocabulary words more heavily once a topic term predicting it is selected is always more effective.

Table 7.12 and Table 7.13 compare the effectiveness of these five sources with QL and SDM as the initial retrieval models respectively. Similarly, Table 7.14 and Table 7.15 presents the same comparison but for the case with RM and CA. We will focus our discussion on the results with PM-2 since it usually achieves better results than xQuAD.

These tables show that, across four models used for retrieving the initial rankings, the terms extracted from Freebase provide the largest performance gain over these baseline rankings in *CPR* and all three cascade diversity measures. These improve-

ments are statistically significant in several cases. They also improve relevance as well, except in the case of CA. We have explained this in the previous section. CA is highly optimized for the relevance of the document rankings using about 100 features. Thus diversifying these rankings based on the query likelihood estimates for $P(d|t)$ (where t is a unigram term) is not ideal. In Chapter 8, we will discuss other types of terms that can help improve these estimates. In addition, the approach based on Freebase outperforms those that use any other resources overall. This confirms the benefits of using a human-edited knowledge base.

AOL, Wiki and RDOC also improve over the initial rankings in almost all measures across QL, SDM and RM. These improvements are statistically significant in some cases (mostly with QL). Regarding *S-Recall*, while the documents obtained from both the retrieval collection and Wikipedia provide no or slight increase in *S-Recall* compared to the initial ranking, AOL (and Freebase) improves *S-Recall* substantially. We believe that this is because the queries and entity descriptions are shorter than the web documents, which makes the extraction of the right topic terms easier. We rule out the reason that the AOL log provides broader coverage for the query topics because we found that the top 50 documents from the initial rankings provide quite broad coverage as well.

Although MSN also provides some improvements, they are not very consistent. The terms generated using MSN help in the case of QL and RM but not in the case of SDM. Our analyses suggest that DSPApprox seems to extract more off-topic terms from MSN than from the other sources, which potentially leads to a larger number of "false topics". As a result, diversification depends more on the documents in the initial rankings (i.e., which "false topics" they happen to match).

Anchor text, on the other hand, is unable to provide improvements over the initial rankings. The reason is that it happens to cover topics that are very different from the ground-truth. For example, the ground-truth topics of the query *"titan"* includes the

"Tennessee Titans" football team and a *"Nissan truck"* model. While most sources provide coverage for these two topics, `Anchor` provides terms such as *"teen"*, *"raven"* and *"cyborg"*. This suggests that `Anchor` covers the topic of *Teen Titan*, which is a comic, and *"raven"* and *"cyborg"* are two of the characters in it. Therefore, `Anchor` indeed provides useful terms. It was penalized simply because our evaluation strategy is based on a predefined topic set. This suggests the potential for combining multiple resources since each of them might cover different topics.

Combining terms (`All`) from all resources does not provide an advantage over using `Freebase` alone with respect to the cascade measures. This is perhaps caused by the terms from `Anchor`. However, `All` consistently improves *S-Recall* across `QL`, `SDM`, `RM` and `CA`. In fact, *S-Recall* provided by `All` is statistically significantly better than both `SDM` and `RM`, which none of the resources individually can achieve. We believe that the reason it is not as good as using `Freebase` alone regarding the three cascade measures is because of the off-topic terms `DSPApprox` provides. Recall that we combine these resources simply by merging top T terms from each of them together. As a result, the number of off-topic terms in the combined approach `All` is quite high.

Note that while `All` is not as effective as `Freebase` regarding the cascade measures and the two relevance measures, it is still substantially better than `QL`, `SDM`, `RM` and slightly better than `CA`. This is very promising considering that the way we combine these resources are very simplistic. We will discuss better methods for combining multiple sources in the last chapter.

Table 7.12: Effectiveness of different sources of information for term extraction. The initial retrieval models are QL. † indicates statistically significant differences (p-value < 0.05) to the initial rankings. Bold face indicates the best performance.

			Diversity							Relevance	
			CPR	α-NDCG	ERR-IA	NRBP	Prec-IA	S-Recall	W/L	NDCG	ERR
QL		Base	0.4422	0.3266	0.217	0.1807	0.1315	0.5289		0.2061	0.1387
	PM-2	RDOC	0.5026†	0.3668†	0.2547†	0.2208†	0.1658†	0.5313	55/28	0.2512†	0.1649†
		AOL	0.5046†	0.3734†	0.2582†	0.2204†	0.1623†	0.5537	53/30	0.2346†	0.1507
		MSN	0.4699	0.3506	0.2363	0.1989	0.1601†	0.552	45/38	0.2176	0.1348
		Anchor	0.4434	0.3182	0.2061	0.1666	0.1436	0.508	37/45	0.1957	0.1313
		Freebase	**0.5137**†	**0.3928**†	**0.2799**†	**0.2495**†	**0.1739**†	**0.5554**	55/29	**0.2678**†	**0.1813**†
		Wiki	0.5088†	0.3735†	0.2655†	0.2357†	0.1676†	0.5207	46/36	0.2547†	0.1779†
		All	0.5066†	0.3654†	0.245†	0.2043	0.1717†	0.5444	50/32	0.2453†	0.1556
	xQuAD	RDOC	0.4935†	0.3639†	**0.2565**†	**0.2259**†	0.166†	0.5259	55/26	0.2523†	0.168†
		AOL	0.4876†	0.3538†	0.2353†	0.1957†	0.1574†	**0.5473**	51/28	0.231†	0.1431
		MSN	0.469	0.3498†	0.2364†	0.1991†	0.1605†	0.544	51/33	0.2221	0.1437
		Anchor	0.4548	0.3309	0.2221	0.1839	0.1436	0.5095	39/42	0.2098	0.1424
		Freebase	0.5054†	**0.3654**†	0.2514†	0.2174†	**0.1744**†	0.541	51/31	**0.2602**†	**0.169**†
		Wiki	**0.5067**†	0.3618†	0.2494†	0.2173†	0.1715†	0.5366	46/35	0.2545†	0.1703†
		All	0.4702†	0.3423	0.2293	0.1902	0.1455	0.5221	46/34	0.2218	0.1432

Table 7.13: Effectiveness of different sources of information for term extraction. The initial retrieval models are SDM. † indicates statistically significant differences (p-value < 0.05) to the initial rankings. Bold face indicates the best performance.

			Diversity							Relevance	
			CPR	α-NDCG	ERR-IA	NRBP	Prec-IA	S-Recall	W/L	NDCG	ERR
SDM		Base	0.4862	0.3593	0.2451	0.2098	0.1777	0.5333		0.2574	0.155
	PM-2	RDOC	0.5079	0.3716	0.2595	0.2267	0.1837	0.5372	47/37	0.2602	0.1686
		AOL	0.5201†	0.3833†	0.2666	0.2291	0.1784	0.5573	47/38	0.2555	0.1601
		MSN	0.4842	0.3565	0.2372	0.1966	0.1805	0.5488	35/47	0.2427	0.1494
		Anchor	0.4841	0.3567	0.2399	0.2002	0.1712	0.5471	46/34	0.2465	0.1527
		Freebase	**0.5222**†	**0.3948**†	**0.2811**†	**0.249**†	0.1968	0.5498	45/39	**0.2862**†	0.1812†
		Wiki	0.5174†	0.3801	0.2669	0.2336	0.1939	0.5347	48/37	0.2815†	**0.1896**†
		All	0.5208†	0.3882†	0.2628	0.223	**0.2009**†	**0.5706**†	51/31	0.2778	0.1766
	xQuAD	RDOC	0.4961	0.3667	0.2569	0.2248	0.187	0.5347	48/33	0.2677	0.172
		AOL	0.4987	0.3718	0.254	0.2136	0.1793	**0.5537**	43/43	0.2523	0.1509
		MSN	0.4696	0.3492	0.2358	0.1975	0.181	0.531	24/29	0.2506	0.1457
		Anchor	0.4849	0.3514	0.2373	0.2001	0.1757	0.524	22/18	0.2563	0.1556
		Freebase	**0.5151**	**0.3856**	**0.2736**	**0.2417**	**0.1961**	0.5359	41/41	**0.29**†	0.182†
		Wiki	0.5066	0.3802	0.2705	0.239	0.1932	0.5398	45/38	0.2822	**0.1878**†
		All	0.4974	0.368	0.2483	0.2102	0.181	0.5515	49/29	0.2579	0.1581

Table 7.14: Effectiveness of different sources of information for term extraction. The initial retrieval models are RM. † indicates statistically significant differences (p-value < 0.05) to the initial rankings. Bold face indicates the best performance.

			Diversity							Relevance	
			CPR	α-NDCG	ERR-IA	NRBP	Prec-IA	S-Recall	W/L	NDCG	ERR
RM		Base	0.4862	0.3486	0.2486	0.2159	0.1776	0.4913		0.2762	0.1668
	PM-2	RDOC	0.5145†	0.3703	0.2606	0.2284	**0.1852**	0.5272†	42/37	0.28	0.1795
		AOL	0.5084	0.3712†	0.2573	0.2213	0.1834	0.5367†	49/30	0.2623	0.1602
		MSN	0.5027	0.3703	0.2588	0.2241	0.1719	0.5398†	43/35	0.2541	0.163
		Anchor	0.4782	0.3509	0.2374	0.2009	0.1648	0.5342†	37/41	0.2377†	0.1459
		Freebase	0.5114	**0.3827**†	**0.2688**	**0.236**	0.183	0.5509†	40/42	0.2816	0.1818
		Wiki	0.5081	0.3663	0.2579	0.2257	0.1839	0.5162	45/36	**0.2828**	**0.1907**
		All	**0.5181**†	0.3794†	0.2608	0.2235	0.184	**0.5515**†	43/34	0.2709	0.1699
	xQuAD	RDOC	0.4929	0.3632	0.2596	0.2292	0.1815	0.5112	42/33	0.2757	0.1709
		AOL	0.505	0.3703	0.2551	0.2166	0.1799	0.5408†	40/42	0.2528	0.151
		MSN	0.4865	0.3562	0.2399	0.2013	0.1757	**0.556**†	41/40	0.2425†	0.1421†
		Anchor	0.4887	0.3503	0.2451	0.2106	0.1747	0.4991	24/32	0.2648	0.1621
		Freebase	**0.5069**	**0.3809**	**0.2725**	**0.243**	0.1828	0.5313	38/42	0.2873	0.1837
		Wiki	0.5023	0.3711	0.2651	0.2353	**0.1849**	0.5216	42/37	**0.2884**	**0.1938**
		All	0.5013	0.3665	0.2535	0.2162	0.1823	0.5262†	44/31	0.2636	0.1631

Table 7.15: Effectiveness of different sources of information for term extraction. The initial retrieval models are CA-rm. † indicates statistically significant differences (p-value < 0.05) to the initial rankings. Bold face indicates the best performance.

			Diversity							Relevance	
			CPR	α-NDCG	ERR-IA	NRBP	Prec-IA	S-Recall	W/L	NDCG	ERR
CA		Base	0.544	0.4159	0.3031	0.2712	0.2149	0.5728		0.3032	0.1971
	PM-2	RDOC	0.5483	0.4077	0.2975	0.2669	0.1991†	0.5524	45/34	0.2939	0.2009
		AOL	0.5451	0.4168	0.3068	0.2774	0.2097	0.5685	45/34	0.3047	0.2048
		MSN	0.542	0.4102	0.2973	0.2672	0.2087	0.5626	46/35	0.3013	0.1973
		Anchor	0.5433	0.4091	0.2974	0.2676	0.209†	0.5587	42/37	0.3001	0.1983
		Freebase	**0.5552**	**0.4212**	**0.3095**	**0.2794**	0.2065†	0.5711	48/34	0.2996	**0.2114**
		Wiki	0.5494	0.4107	0.2961	0.2635	0.2084	0.5731	46/37	0.2996	0.1964
		All	0.5485	0.418	0.3043	0.2735	0.2098	**0.5757**	43/39	**0.3045**	0.1998
	xQuAD	RDOC	0.5428	0.4113	0.2993	0.2687	0.2106†	0.5677	46/34	**0.3022**	0.1966
		AOL	0.5418	0.4074	0.2963	0.2649	0.2099†	0.5643	45/36	0.2988	0.1942
		MSN	0.5422	0.4095	0.2978	0.2667	0.2103†	0.5677	48/33	0.2993	0.1943
		Anchor	**0.5447**	**0.4168**	**0.3052**	**0.2758**	0.2084†	**0.5745**	47/35	0.2992	**0.1996**
		Freebase	0.5431	0.4099	0.298	0.267	0.2107†	0.5677	49/32	0.3001	0.1947
		Wiki	0.5421	0.409	0.2966	0.2652	0.211†	0.5677	49/32	0.3	0.1942
		All	0.538	0.4007	0.2854	0.2534	0.2082†	0.566	41/40	0.2988	0.192

7.5 Summary

In this chapter, we have demonstrated the effectiveness of our term level approach to diversification. We show, using both the TREC sub-topics and the related queries from a commercial search engine, that diversifying a result ranking with respect to a set of topic terms can promote diversity with respect to the topics underlying these terms. In other words, term level diversification can be considered diversification with respect to the underlying latent topics.

This effectively reduces the task of finding a set of query topics, which has proven difficult, to finding a set of terms. Consequently, we show that such topic terms can be extracted automatically and effectively using `DSPApprox` (Lawrie & Croft, 2003), a technique previously proposed for document summarization.

The mechanism used by `DSPApprox` reveals that generating topic terms is, in fact, a diversification problem by itself. Although this suggests that existing techniques for document diversification such as `PM-2` and `xQuAD` can be applied, this turns out not to be the case, at least not directly. The reason is because these techniques typically downweight a topic as a document on this topic is selected. In the context of term generation where "topics" correspond to vocabulary and "documents" correspond to topic terms, there are thousands of vocabulary words for which a result set of topic terms should provide coverage. This lenient nature of downweighting is not sufficient to ensure the next topic term to be selected will cover a different part of the vocabulary. On the other hand, the aggressive nature of `DSPApprox`, which completely ignores all vocabulary as soon as a topic term covering them is selected, has been shown to be very successful.

Finally, we show that documents in the initial rankings, query logs, anchor text, Freebase and Wikipedia pages are valuable sources for extracting topic terms. Each of them individually can provide effective terms for diversification. Among these, Free-

base provides the largest performance gain in both diversity and relevance measures. Combining them further improves topic coverage in the search results.

CHAPTER 8

CONCLUSIONS AND FUTURE WORK

8.1 Conclusions

In this thesis, we introduced a new perspective to search result diversification: diversity by proportionality. Instead of quantifying diversity by the amount of novelty in a result ranking, we consider a list more diverse if the ratio between the number of documents it provides for each query topic matches more closely with the topic popularity distribution. Based on this perspective, we derived an effectiveness measure called Cumulative Proportionality (*CPR*) and a framework for optimizing proportionality in search results with two instantiations: `PM-1` and `PM-2`. While `PM-1` is a simple adaptation of the Sainte-Laguë method used for promoting proportionality in elections, it serves as a basis for `PM-2`, a practical adaptation that takes into account the fact that a document might be related to multiple topics.

Regarding our *CPR* measure, we found that it correlates well with existing diversity measures that are based on novelty and redundancy. The reason is that enforcing proportionality at every ranks helps surface relevant documents for more topics at early positions in the list, thereby increasing novelty and equivalently decreasing redundancy.

Despite the correlation, *CPR* is different to those measures in the following way. The effectiveness of a diverse ranking comprises of three factors: the number of topics for which it can provide at least one relevant document, the number of relevant documents it contains for each topic (which implies the number of non-relevant documents) and the positions of these relevant documents in this ranking. Among the

rankings that are only effective with respect to one or two out of the three criteria, existing measures prefer those in which the relevant documents (to any of the topics) are highly ranked even though the overall topic coverage is low and there are only a few such relevant documents. *CPR*, on the other hand, favors the result lists in which the relevant documents might not appear in the highest positions but there is coverage for a larger number of topics as well as more relevant documents per topic (which reflects higher proportionality to some extent).

We demonstrated, using the set of ground-truth topics as well as the related queries obtained from a commercial search engine, that `PM-2` outperforms `xQuAD`, a top performing redundancy-based diversification technique, with respect to both *CPR* and a variety of redundancy-based diversity measures which `xQuAD` was designed to optimize. The improvement provided by `PM-2` is consistent across four standard baseline retrieval models – Query Likelihood (Ponte & Croft, 1998), Sequential Dependence Model (Metzler & Croft, 2005), Relevance Model (Lavrenko & Croft, 2001) and a learning to rank model trained using Coordinate Ascent (Metzler & Croft, 2007).

Recall that `xQuAD` measures the novelty of a document based on the marginal probability that it covers the topics with low probability of having been covered by those selected earlier. `PM-2`, on the other hand, scores a document based on how proportional the resulting set of documents becomes if this document is selected, which loosely models its novelty. As a result, `PM-2` can be considered a redundancy minimization technique. Our analyses have shown that this proportionality-based novelty measure is one of the factors that makes `PM-2` better.

Since diversification is not practical without the ability to automatically infer the topics associated with the user queries, we address the problem of topic generation. We have shown that reformulations for queries obtained from two publicly available resources, anchor text and Microsoft Web N-Gram Services, can provide coverage for a broad range of query topics. We experimented with two clustering algorithms

and two query similarity measures for grouping the reformulations that are topically related. We found that agglomerative clustering using the similarity measure based on co-occurence provides the most topically consistent clusters. Furthermore, our retrieval experiments have shown that these clusters, each of which is assumed to represent a query topic, are also effective for diversification.

We observed that modeling for each query a set of topics where each topic is a group of terms makes the task of topic generation very difficult. Consequently, we proposed to identify for each query a set of terms that describe these topics. These terms are then provided to existing diversification techniques that treat each of them as a topic for which they aim to provide coverage. This approach is called term level diversification. Our hypothesis is that diversification using a set of terms can promote diversity with respect to the latent topics underlying these terms, and thus being able to identify the important topic terms related to the true query topics is sufficient for improving diversity in a result ranking. This was confirmed by our retrieval experiments using the set ground-truth topics and the corresponding set of terms as well as the related queries from a commercial search engine.

Our term level approach to diversification effectively reduces the task of finding a set of query topics, which has proven difficult, into finding a set of topic terms. Consequently, we used `DSPApprox`, a greedy algorithm from the literature of multi-document summarization (Lawrie & Croft, 2003) to identify a diverse set of terms (unigrams and phrases). Our results indicated that the term level approach using these terms achieves significantly better results compared to the topic level counterpart using topics that are generated automatically by two existing cluster-based techniques.

Interestingly, we showed that `DSPApprox` is itself a diversification algorithm. Although this suggests that existing techniques for document diversification such as `PM-2` and `xQuAD` can be used to generate effective topic terms, this turned out not to

be the case, at least not directly. Our analyses of the similarity between DSPApprox and these methods demonstrated that document diversification techniques are too lenient in penalizing redundancy. They typically downweight a topic each time a document on this topic is selected. In the context of term generation where "topics" correspond to vocabulary and "documents" correspond to topic terms, there are thousands of vocabulary words for which a result set of topic terms should provide coverage. This lenient downweighting is not sufficient to ensure the next topic term to be selected will cover a different part of the vocabulary. On the other hand, the aggressive nature of DSPApprox, which completely ignores a subset of the vocabulary as soon as a topic term covering them is selected, has been shown to be very successful.

Finally, we explored several sources of information for generating topic terms. This includes documents in the initial rankings, query logs, anchor text, Freebase and Wikipedia pages. We have shown that each of them individually can provide effective terms for diversification. Among these resources, Freebase provides the largest performance gain in both diversity and relevance measures. Combining them helps lead to broader topic coverage.

8.2 Future Work

As mentioned before, the failure with PM-2 that we observed is due the fact that we use Query Likelihood to estimate the relevance between a candidate document to the query topics ($P(d|t)$). We intend to apply the learning to rank framework to estimate $P(d|t)$. We believe that this will substantially improve both the diversity and relevance in the final ranking.

An interesting direction is to extend PM-2 to work in the multi-level diversification setting. Ideally, if the result list for the query "java" should contain nine documents related to the programming language, we should consider diversifying these nine documents as well to account for different sub-topics within programming.

Recall that our diversification experiments are conducted on top of the initial rankings retrieved using different relevance-based models. Among these, the initial rankings that are obtained using the learning to rank approach outperforms all other non-learning baselines (Query Likelihood, Sequential Dependency and Relevance Model) for all diversity measures. It is important to point out that the learning to rank approach, in fact, does not promote diversity. It is optimized to retrieve more relevant results, which naturally increases the chance of covering more topics, thereby improving diversity.

Yue and Joachims (2008) propose a learning framework that truly optimizes for diversity. As opposed to the learning to rank approach which takes as input a document and outputs its relevance estimate, this framework aims to learn a function $f_w(S, R) = w^T \Phi(S, R)$ which takes as input a set of document S and outputs its coverage with respect to the set of topics in target document set R. The feature set $\Phi(S, R)$ describes how well S covers the vocabulary in R. The model parameter w is learned from training data using structural SVM. At run time, this model is applied in the same greedy fashion. It iteratively selects a document from the input ranking R to put into the output (diverse) ranking S. At each iteration, it selects the candidate documents d with maximum $f_w(S \cup d, R)$.

Although Yue and Joachims (2008) focus on the coverage of the result ranking, it is possible to incorporate other features. The limitation of this framework, on the other hand, is that of the implicit approach: it does not model query topics, thus cannot take into account topic popularity. We will explore the possibility of integrating proportionality into this framework in the future. The benefits of a learning framework is that we can incorporated arbitrarily defined features, such as the topics generated from multiple sources. However, this poses a major challenge to the notion of proportionality, which by definition is with respect to a single set of query topics. If one can clearly define what proportionality means in this case, this potentially helps

with multi-dimension diversification well (i.e., the search results provide coverage for not for only multiple topics but also for multiple sentiments per topic).

Although it is unclear how this learning framework could be used to optimize for proportionality, it can be used to generate topic terms. `DSPApprox`, `PM-2` and `xQuAD` all employ their own heuristic combination of topicality and predictiveness. These two features can be incorporated into the feature set $\Phi(S, R)$ easily where S is now the set of topic terms. Any other potentially useful features could be added as well. In fact, predictiveness itself can be modeled using multiple features, each of which corresponds to a different subset of the vocabulary in R. For example, we could have one feature indicating how well S predicts the entire set of vocabulary and another indicating how well S predicts the set of vocabulary that occurs at least five times in R. In addition, the co-occurrence statistics between a topic term and a vocabulary word estimated from multiple sources of information could be added as separate features. The idea is that a topic term that consistently predicts a set of vocabulary words across different data sources should be a good topic term. The burstiness of a term can potentially be a useful as well since terms that occur very frequently only in a few documents should be highly topical. A major challenge is how we construct the training data. We can start with treating the terms from the ground-truth topics as the optimal set of topic terms.

Regarding the types of terms used for diversification, we have experimented with unigrams and phrases. As shown in Chapter 4, as we use more effective models to retrieve the initial ranking for diversification, phrases usually perform slightly better than unigrams for all three cascade measures as well as the traditional relevance measures. This is very promising given that the method we used to extract phrases are very simplistic. In the future, we want to study whether using more sophisticated NLP techniques for phrase extraction can further improve diversity effectiveness.

In addition to unigrams and phrases, we plan to consider anchor text as a "term". Recall that the basis of our term level approach is that existing diversification techniques can take as input a set of terms and provide a diverse ranking with respect to the latent topics underlying these terms. This is because these techniques favor documents that are "relevant" to multiple terms. Intuitively, a latent topic will have a higher chance of being covered in the result ranking if it is represented by a sufficient number of terms in the input set. Our results in Chapter 5 have shown that we can generate topically consistent clusters of anchor text. This indicates that we can obtain a sufficient number of anchor texts for each topic. It would be interesting to see if disregarding the cluster structure and using all of the reformulations as a set of terms for diversification can improve diversity. Furthermore, how is this approach compared to using the clusters for diversification as we have done earlier? Each cluster likely has outliers, which might affect the estimate of $P(d|t)$. We suspect that not using the cluster structures is more effective (which will further demonstrate the effectiveness of our term level approach) since it is less likely that there is a sufficient number of outliers that consistently represent some outlying topic. This might raise the question why we need clustering in the first place. We suspect that clustering is necessary since it helps filter out as many unrelated reformulations as possible. Furthermore, the clusters might provide a basis to estimate the popularity of the query topics.

Note that we can also consider the queries in search logs and the entities in Freebase as terms. Compared to using unigrams, using phrases, anchor text and entities as terms have the advantage of containing more text, which helps the $P(d|t)$ estimate (i.e., one can use proximity features which are not available with unigram terms).

At the moment, the topic terms provided to both `PM-2` and `xQuAD` for diversification are assumed to have equal weights. This is because it is unclear how the weight of these terms reflects the popularity of the underlying topics. In the future, we will

experiment with weighting these topic terms by their frequency estimated from some collection (e.g., retrieval collection, query logs, Wikipedia, etc.). Comparing the ratio between the number of relevant documents returned for each of the query topics in this case to the case with uniform term weighting might shed some light on how term weighting reflects the topic popularity.

REFERENCES

Agichtein, E., Brill, E., & Dumais, S. (2006). Improving web search ranking by incorporating user behavior information. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 19–26).

Agrawal, R., Gollapudi, S., Halverson, A., & Ieong, S. (2009). Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining* (pp. 5–14).

Aktolga, E. (2014). *Integrating non-topical aspects into information retrieval* (Doctoral Dissertation). University of Massachusetts Amherst.

Aktolga, E., & Allan, J. (2013). Sentiment diversification with different biases. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 593–602).

Allan, J., Carbonell, J., Doddington, G., Yamron, J., & Yang, Y. (1998, February). Topic detection and tracking pilot study: Final report. In *DARPA Broadcast News Transcription and Understanding Workshop* (pp. 194–218). Lansdowne, VA, USA.

Allan, J., Gupta, R., & Khandelwal, V. (2001). Temporal summaries of new topics. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 10–18).

Allan, J., Papka, R., & Lavrenko, V. (1998). On-line new event detection and tracking. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 37–45).

Allan, J., Wade, C., & Bolivar, A. (2003). Retrieval and novelty detection at the sentence level. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 314–321).

Ashkan, A., & Clarke, C. L. (2011). On the informativeness of cascade and intent-aware effectiveness measures. In *Proceedings of the 20th International Conference on World Wide Web* (pp. 407–416).

Bendersky, M., Croft, W. B., & Diao, Y. (2011). Quality-biased ranking of web documents. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining* (pp. 95–104).

Bendersky, M., Fisher, D., & Croft, W. B. (2010). UMass at TREC 2010 web track: Term dependence, spam filtering and quality bias. In *Proceedings of the 19th Text REtrieval Conference.*

Bendersky, M., Metzler, D., & Croft, W. B. (2012). Effective query formulation with multiple information sources. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining* (pp. 443–452).

Berberich, K., & Bedathur, S. (2004). Temporal diversification of search results. In *Proceedings of Workshop on Time-aware Information Access (the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval).*

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993-1022.

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., & Taylor, J. (2008). Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data* (pp. 1247–1250).

Buckley, C., & Voorhees, E. M. (2004). Retrieval evaluation with incomplete information. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 25–32).

Burges, C. J., Ragno, R., & Le, Q. V. (2006). Learning to rank with nonsmooth cost functions. In *Advances in Neural Information Processing Systems* (pp. 193–200).

Burges, C. J., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., & Hullender, G. N. (2005). Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine Learning* (pp. 89–96).

Cai, D., Mei, Q., Han, J., & Zhai, C. (2008). Modeling hidden topics on document manifold. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management* (pp. 911–920).

Carbonell, J., & Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 335–336).

Carterette, B., & Chandar, P. (2009). Probabilistic models of ranking novel documents for faceted topic retrieval. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management* (pp. 1287–1296).

Chapelle, O., Metlzer, D., Zhang, Y., & Grinspan, P. (2009). Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management* (pp. 621–630).

Chen, H., & Karger, D. R. (2006). Less is more: Probabilistic models for retrieving fewer relevant documents. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 429–436).

Clarke, C. L., & Craswell, N. (2012). Overview of the TREC 2012 web track. In *Proceedings of the 21th Text REtrieval Conference.*

Clarke, C. L., Craswell, N., & Soboroff, I. (2009). Overview of the TREC 2009 web track. In *Proceedings of the 18th Text REtrieval Conference.*

Clarke, C. L., Craswell, N., Soboroff, I., & Ashkan, A. (2011). A comparative analysis of cascade measures for novelty and diversity. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining* (pp. 75–84).

Clarke, C. L., Craswell, N., Soboroff, I., & Cormack, G. V. (2010). Overview of the TREC 2010 web track. In *Proceedings of the 19th Text REtrieval Conference.*

Clarke, C. L., Craswell, N., Soboroff, I., & Voorhees, E. M. (2011). Overview of the TREC 2011 web track. In *Proceedings of the 20th Text REtrieval Conference.*

Clarke, C. L., Kolla, M., Cormack, G. V., Vechtomova, O., Ashkan, A., Büttcher, S., & MacKinnon, I. (2008). Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 659–666).

Clarke, C. L., Kolla, M., & Vechtomova, O. (2009). An effectiveness measure for ambiguous and underspecified queries. In *Proceedings of the 2nd International Conference on Theory of Information Retrieval: Advances in Information Retrieval Theory* (pp. 188–199).

Clough, P., Sanderson, M., Abouammoh, M., Navarro, S., & Paramita, M. (2009). Multiple approaches to analysing query diversity. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 734–735).

Cormack, G. V., Smucker, M. D., & Clarke, C. L. (2011). Efficient and effective spam filtering and re-ranking for large web datasets. *Information Retrieval*, *14*(5).

Craswell, N., Zoeter, O., Taylor, M., & Ramsey, B. (2008). An experimental comparison of click position-bias models. In *Proceedings of the 2008 International Conference on Web Search and Data Mining* (pp. 87–94).

Croft, W. B., Metzler, D., & Strohman, T. (2009). *Search engines: Information retrieval in practice.* Addison-Wesley.

Cronen-Townsend, S., Zhou, Y., & Croft, W. B. (2002). Predicting query performance. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 299–306).

Dang, V., Bendersky, M., & Croft, W. B. (2013). Two-stage learning to rank for information retrieval. In *Proceedings of the 35th European Conference on Advances in Information Retrieval* (pp. 423–434).

Dang, V., & Croft, W. B. (2010). Query reformulation using anchor text. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining* (pp. 41–50).

Dang, V., & Croft, W. B. (2013). Term level search result diversification. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 603–612).

Demartini, G. (2011). ARES: A retrieval engine based on sentiments sentiment-based search result annotation and diversification. In *Proceedings of the 33rd European Conference on Advances in Information Retrieval* (pp. 772–775).

Dou, Z., Hu, S., Chen, K., Song, R., & Wen, J.-R. (2011). Multi-dimensional search result diversification. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining* (pp. 475–484).

Eiron, N., & McCurley, K. S. (2003). Analysis of anchor text for web search. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 459–460).

Feige, U. (1998, July). A threshold of ln n for approximating set cover. *Journal of the ACM*, *45*(4), 634–652.

Freund, Y., Iyer, R. D., Schapire, R. E., & Singer, Y. (2003). An efficient boosting

algorithm for combining preferences. *Journal of Machine Learning Research*, *4*, 933-969.

Fuxman, A., Tsaparas, P., Achan, K., & Agrawal, R. (2008). Using the wisdom of the crowds for keyword generation. In *Proceedings of the 17th International Conference on World Wide Web* (pp. 61–70).

Gallagher, M. (1991). Proportionality, disproportionality and electoral systems. *Electoral Studies*, *10*(1), 33–51.

Goffman, W. (1964). On relevance as a measure. *Information Storage and Retrieval*, *2*(3), 201–203.

Harman, D. (2002). Overview of the TREC 2002 novelty track. In *Proceedings of the 11th Text REtrieval Conference.*

He, J., Hollink, V., & de Vries, A. (2012). Combining implicit and explicit topic representations for result diversification. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 851–860).

Hersh, W., & Over, P. (1999). TREC-8 interactive track report. In *Proceedings of the 8th Text REtrieval Conference.*

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

Jones, R., & Diaz, F. (2007). Temporal profiles of queries. *ACM Transactions on Information Systems*, *25*(3).

Jones, R., Rey, B., Madani, O., & Greiner, W. (2006). Generating query substitutions. In *Proceedings of the 15th International Conference on World Wide Web* (pp. 387–396).

Kacimi, M., & Gamper, J. (2011). Diversifying search results of controversial queries. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management* (pp. 93–98).

Kanungo, T., & Orr, D. (2009). Predicting the readability of short web summaries. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining* (pp. 202–211).

Keikha, M., Crestani, F., & Croft, W. B. (2012). Diversity in blog feed retrieval. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management* (pp. 525–534).

Krovetz, R. (1993). Viewing morphology as an inference process. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 191–202).

Larkey, L., Allan, J., Connell, M., Bolivar, A., & Wade, C. (2002). Umass at trec 2002: Cross language and novelty tracks. In *Proceedings of the 11th Text REtrieval Conference.*

Lavrenko, V., & Croft, W. B. (2001). Relevance based language models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 120–127).

Lawrie, D., & Croft, W. B. (2003). Generating hierarchical summaries for web searches. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 457–458).

Lawrie, D., Croft, W. B., & Rosenberg, A. (2001). Finding topic words for hierarchical summarization. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 349–357).

Liu, T.-Y. (2009). Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, *3*(3), 225–331.

Ma, H., Lyu, M. R., & King, I. (2010). Diversifying query suggestion results. In *Association for the Advancement of Artificial Intelligence.*

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, *7*(1), 77–91.

McCreadie, R., Macdonald, C., Santos, R. L. T., & Ounis, I. (2011). University of Glasgow at TREC 2011: Experiments with Terrier in crowdsourcing, microblog, and web tracks. In *Proceedings of the 20th Text REtrieval Conference.*

Metzler, D., & Croft, W. B. (2005). A Markov random field model for term dependencies. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 472–479).

Metzler, D., & Croft, W. B. (2007, June). Linear feature-based models for information retrieval. *Information Retrieval*, *10*(3), 257–274.

Metzler, D., Novak, J., Cui, H., & Reddy, S. (2009). Building enriched document representations using aggregated anchor text. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 219–226).

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

Nemhauser, G., Wolsey, L., & Fisher, M. (1978). An analysis of approximations for maximizing submodular set functions–I. *Mathematical Programming*, *14*(1), 265-294.

Ntoulas, A., Najork, M., Manasse, M., & Fetterly, D. (2006). Detecting spam web pages through content analysis. In *Proceedings of the 15th International Conference on World Wide Web* (pp. 83–92).

Over, P. (1997). TREC-6 interactive track report. In *Proceedings of the 6th Text REtrieval Conference.*

Over, P. (1998). TREC-7 interactive track report. In *Proceedings of the 7th Text REtrieval Conference.*

Ponte, J. M., & Croft, W. B. (1998). A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 275–281).

Radlinski, F., Kleinberg, R., & Joachims, T. (2008). Learning diverse rankings with multi-armed bandits. In *Proceedings of the 25th International Conference on Machine Learning* (pp. 784–791).

Radlinski, F., Szummer, M., & Craswell, N. (2010). Inferring query intent from reformulations and clicks. In *Proceedings of the 19th International Conference on World Wide Web* (pp. 1171–1172).

Rafiei, D., Bharat, K., & Shukla, A. (2010). Diversifying web search results. In *Proceedings of the 19th International Conference on World Wide Web* (pp. 781–

790).

Raman, K., Joachims, T., & Shivaswamy, P. (2011). Structured learning of two-level dynamic rankings. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management* (pp. 291–296).

Raman, K., Shivaswamy, P., & Joachims, T. (2012). Online learning to diversify from implicit feedback. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 705–713).

Robertson, S. E. (1997). The probability ranking principle in IR. In *Readings in Information Retrieval* (pp. 281–286).

Robertson, S. E., & Walker, S. (1994). Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 232–241).

Sakai, T., & Song, R. (2011). Evaluating diversified search results using per-intent graded relevance. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 1043–1052).

Sanderson, M., & Croft, W. B. (1999). Deriving concept hierarchies from text. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 206–213).

Sanderson, M., Paramita, M. L., Clough, P., & Kanoulas, E. (2010). Do user preferences and evaluation measures line up? In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 555–562).

Santos, R. L., Macdonald, C., & Ounis, I. (2010a). Exploiting query reformulations for web search result diversification. In *Proceedings of the 19th International Conference on World Wide Web* (pp. 881–890).

Santos, R. L., Macdonald, C., & Ounis, I. (2010b). Selectively diversifying web search results. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management* (pp. 1179–1188).

Santos, R. L., Macdonald, C., & Ounis, I. (2011). Intent-aware search result diversification. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 595–604).

Santos, R. L., Macdonald, C., & Ounis, I. (2013, August). Learning to rank query suggestions for adhoc and diversity search. *Information Retrieval*, *16*(4), 429–451.

Simpson, E. (1949). Measurement of diversity. *Nature*, *163*(4148), 688.

Slivkins, A., Radlinski, F., & Gollapudi, S. (2010). Learning optimally diverse rankings over large document collections. In *Proceedings of the 27th International Conference on Machine Learning* (pp. 983–990).

Soboroff, I. (2004). Overview of the TREC 2004 novelty track. In *Proceedings of the 13th Text REtrieval Conference.*

Soboroff, I., & Harman, D. (2003). Overview of the TREC 2003 novelty track. In *Proceedings of the 12th Text REtrieval Conference.*

Song, R., Luo, Z., Wen, J.-R., Yu, Y., & Hon, H.-W. (2007). Identifying ambiguous queries in web search. In *Proceedings of the 16th International Conference on*

World Wide Web (pp. 1169–1170).

Song, Y., Zhou, D., & He, L.-w. (2011). Post-ranking query suggestion by diversifying search results. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 815–824).

Wang, J., & Zhu, J. (2009). Portfolio theory of information retrieval. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 115–122).

Wu, Q., Burges, C. J., Svore, K., & Gao, J. (2010). Adapting boosting for information retrieval measures. *Information Retrieval*, *13*(3).

Xu, J., & Li, H. (2007). Adarank: A boosting algorithm for information retrieval. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 391–398).

Xu, Y., Jones, G. J., & Wang, B. (2009). Query dependent pseudo-relevance feedback based on Wikipedia. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 59–66).

Yue, Y., & Guestrin, C. (2011). Linear submodular bandits and their application to diversified retrieval. In *Advances in Neural Information Processing Systems* (pp. 2483–2491).

Yue, Y., & Joachims, T. (2008). Predicting diverse subsets using structural SVMs. In *Proceedings of the 25th International Conference on Machine Learning* (pp. 1224–1231).

Zhai, C. X., Cohen, W. W., & Lafferty, J. (2003). Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 10–17).

Zheng, W., Wang, X., Fang, H., & Cheng, H. (2011). An exploration of pattern-based subtopic modeling for search result diversification. In *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries* (pp. 387–388).

Zheng, W., Wang, X., Fang, H., & Cheng, H. (2012, October). Coverage-based search result diversification. *Information Retrieval*, *15*(5), 433–457.